UNIVERSITY OF CALIFORNIA
Los Angeles

The sound pattern of Japanese surnames

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Linguistics

by

Yu Tanaka

2017

ABSTRACT OF THE DISSERTATION

The sound pattern of Japanese surnames

by

Yu Tanaka
Doctor of Philosophy in Linguistics
University of California, Los Angeles, 2017
Professor Kie Ross Zuraw, Co-chair
Professor Bruce P. Hayes, Co-chair

Compound surnames in Japanese show complex phonological patterns, which pose challenges to current theories of phonology. This dissertation proposes an account of the segmental and prosodic issues in Japanese surnames and discusses their theoretical implications.

Like regular compound words, compound surnames may undergo a sound alternation known as *rendaku*, whereby the initial consonant of the second element becomes voiced (e.g. /yama + t̲a/ → [yama-**d**a] 'mountain-paddy'). The voicing alternation in surnames is somewhat different from that in regular compounds, however; its application is often affected by the features of the last consonant of the first element. Surnames also show unique prosodic patterns, which include an inverse correlation between accentedness and rendaku application. Although the peculiarities of surnames have been noted in the literature (see Sugito 1965 among others), no study has ever provided a full description or explanation of the patterns. The first goal of the dissertation is to account for why compound surnames are different from regular compounds in terms of rendaku application and accentuation.

I claim that compound surnames are represented as single stems in the grammar due to their semantic non-compositionality and that their peculiar phonological patterns can be attributed to the application of stem-internal phonology. I present the results of a corpus study of existing surnames collected from social media, and a rendaku judgment experiment using nonce surnames.

Both studies support the hypothesis that compound surnames follow the phonology of stems. The analysis opens a new way to investigate the sound patterns of proper nouns in general.

Rendaku in surnames poses another theoretical problem since it exhibits both lexical irregularities and phonologically-conditioned productivity. The experimental results show that Japanese speakers apply rendaku productively based on phonological factors. However, a closer look at the patterns of real surnames indicates that rendaku is also highly lexicalized; besides phonological factors, the presence of voicing in a given surname is determined by the idiosyncratic properties of that surname. The challenge of capturing the lexicalized and productive aspects of a phonological phenomenon with a single grammar has been recognized but not always addressed in the literature (see Zuraw 2000; Moore-Cantwell and Pater 2016).

To meet this challenge, I propose a Maximum Entropy Harmonic Grammar model (see Goldwater and Johnson 2003) with general phonological constraints and lexically-specific constraints along the lines of Moore-Cantwell and Pater (2016). I show that the proposed model with appropriate biases on the learning of constraint weights can not only capture the lexicalized rendaku patterns of existing surnames but also predict productive rendaku application in non-existing surnames. The analysis suggests that lexical factors should be incorporated into phonological grammar rather than simply specified in the lexicon.

The dissertation of Yu Tanaka is approved.

Junko Ito

Megha Sundara

Bruce P. Hayes, Committee Co-chair

Kie Ross Zuraw, Committee Co-chair

University of California, Los Angeles

2017

いろはにほへと ちりぬるを
わかよたれそ つねならむ
うゐのおくやま けふこえて
あさきゆめみし ゑひもせす

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

ACKNOWLEDGMENTS

My dissertation project has been supported by many people. First and foremost, I would like to express my deep gratitude to my committee chairs Kie Zuraw and Bruce Hayes. As everybody in the UCLA Linguistics Department knows, they are amazing advisors. Without their support, this dissertation would not have existed. In regular meetings with Kie, I have had countless fruitful discussions. Indeed, that's where many of the ideas presented in this work were born. I also just had so much fun talking about language and issues in linguistics with her. Bruce always asked difficult questions. I could not answer all of them, but just trying to answer them helped me better understand the problems I was tackling. I also very much enjoyed his intellectual jokes, even though they were sometimes difficult for me to understand. I will definitely miss those meetings with Kie and Bruce.

I'm also grateful to the other members of my committee. Megha Sundara taught me how to run experiments and analyze results from the basics. Discussions with her were very helpful in improving the design of the experiment for this research. As an external member, Junko Ito gave me a lot of insightful comments and suggestions, which helped me develop the formal analyses presented in this dissertation. I should also add that her works with Armin Mester on various aspects of Japanese phonology have influenced me as a linguist in a lot of ways. It was a great honor to work with her.

The following people also gave me helpful comments on this research project: Atsushi Asai, Adam Chong, Robert Daland, Eleanor Glewwe, Shigeto Kawahara, Kazutaka Kurisu, Armin Mester, Eric Rosen, Shin-ichiro Sano, Brian Smith, Ayaka Sugawara, Wataru Uegaki, Timothy Vance and Meng Yang. I also thank the audiences at LSA 2017, AMP 2016, WCCFL 34, WAFL 12 and FAJL 8 as well as the members of the UCLA Phonetics/Phonology Seminars for their feedback.

I'm indebted to the faculty members of the UCLA Linguistics Department for the wonderful academic training they've given me. I extend my gratitude especially to Robert Daland, Bruce Hayes, Sun-Ah Jun, Patricia Keating, Russel Schuh, Megha Sundara and Kie Zuraw. Many

thanks also go to my fellow graduate students at UCLA for having provided me with delightful academic and non-academic experiences: Natasha Abner, Jason Bishop, Dustin Bowers, Margit Bowler, Adam Chong, Philippe Côté-Boucher, Elyssa Dudley, Marc Garellek, Eleanor Glewwe, John Gluckman, Vanya Kapitonov, Natasha Korotkova, Michael Lefkowitz, Isabelle Lin, Nicoletta Loccioni, Laura McPherson, Sadie Martin, Travis Major, Iara Mantenuto, Victoria Mateu, Chase O'Flynn, Deniz Özyıldız, Richard Stockwell, Chad Vicenik, David Wemhaner, Jamie White, Meng Yang, Jun Yashima, Jesse Zymet and many others.

I thank Yoriko Kashihara, Ryoichiro Kobayashi, Kazuko Kondo, Eri Osawa, Ayaka Sugawara, Shin-ichi Tanaka, Yayoi Tanaka, Kanako Tomaru and Tomoyuki Yoshida for their help in participant recruitment and data collection. In addition, I thank the participants of the experiment who provided the data of this study. Thanks to Mark Irwin for letting me use data from his rendaku corpus presented in Chapter 4.

I also want to express my special thanks to Shigeko Shinohara, Takahito Shinya, Donna Erickson, Naoki Fukui, Gen Fujita and my other former colleagues at Sophia University. Without them, I would not have even thought about pursuing a Ph.D. in linguistics.

I'm also grateful to my parents Michiko and Shinobu Tanaka for their constant support throughout my graduate life. My thanks also go to my brother Ken Akutsu, my sister-in-law Rie and my little niece and nephew Utako and Chiaki. Going back to Japan and playing with them once a year was a nice break from work.

There are still many other people who I want to acknowledge here, but unfortunately space limitations do not allow me to do so. UCLA, Los Angeles, the people I met here and those I met in lots of other places I visited in the last five and a half years have treated me so well. I really appreciate your support. Arigato!

VITA

2002–2007	Meiji University, Tokyo, Japan B.A. in French Literature. *Summa Cum Laude*.
2009–2011	Sophia University, Tokyo, Japan. M.A. in Linguistics.
2011–2017	University of California, Los Angeles Ph.D. in Linguistics (expected)

PUBLICATIONS AND PRESENTATIONS

Tanaka, Yu. (2017) Phonotactically-driven rendaku in surnames: A linguistic study using social media. *Proceedings of the West Coast Conference on Formal Linguistics (WCCFL) 34*. 519–528.

Tanaka, Yu (2015) The perceptual basis of the skewed distributions of Japanese palatalized consonants. *Proceedings of the North East Linguistic Society (NELS) 45.* Vol 3: 131–144.

Tanaka, Yu (2014) The role of contrast in the distributional restrictions on Japanese palatalized consonants. *Proceedings of Formal Approaches to Japanese Linguistics (FAJL) 7*: 239–250.

Tanaka, Yu and Jun Yashima (2013) Deliberate Markedness in Japanese hypocoristics. *Proceedings of GLOW in Asia IX*: 283–297.

Tanaka, Yu (2017) Modeling productive rendaku application in real and nonce Japanese surnames. *The 91st Annual Meeting of the Linguistic Society of America (LSA 2017)*. Austin, TX, USA. January 5th-8th, 2017.

Tanaka, Yu (2016) Implosives in Jakarta Indonesian. *The 5th Joint meeting of the Acoustical Society of America and the Acoustical Society of Japan*. Honolulu, HI, USA. November 28th-December 2nd, 2016

Tanaka, Yu (2016) The correlation between rendaku and accent in Japanese surnames: A foot-based account. *The 2016 Annual Meeting on Phonology (AMP 2016)*. University of Sourthern California. October 21st-23rd, 2016.

Tanaka, Yu (2016) Monomorphemicity of proper names: Evidence from rendaku in Japanese surnames. *The 12th Workshop on Altaic Formal Linguistics (WAFL 12)*. Central Connecticut State University. May 12th-15th, 2016.

Tanaka, Yu (2016) A linguistic study using social media: Phonotactically-driven rendaku in surnames. *The 34th West Coast Conference on Formal Linguistics (WCCFL 34)*. University of Utah. April 29-May 1st, 2016.

Tanaka, Yu (2016) Rendaku in Japanese surnames revisited: Just pretending to be monomorphemic! *The 8th Meeting of Formal Approaches to Japanese Linguistics (FAJL 8)*. Mie University. February 18th-20th, 2016.

Tanaka, Yu (2014) The perceptual basis of the skewed distributions of Japanese palatalized consonants. *Palatalization*. University of Tromsø. December 4th-5th, 2014.

Tanaka, Yu (2014) The perceptual basis of the skewed distributions of Japanese palatalized consonants. *North East Linguistic Society (NELS) 45*. Massachusetts Institute of Technology. October 31st-November 2nd, 2014.

Tanaka, Yu (2014) The role of contrast in the distributional restrictions on Japanese palatalized consonants. *Formal Approaches to Japanese Linguistics (FAJL) 7*. NINJAL/ICU. June 27th-29th, 2014.

Tanaka, Yu and Jun Yashima (2012) Deliberate Markedness in Japanese hypocoristics. *GLOW in Asia IX*. Mie University. September 4th-6th, 2012.

CHAPTER 1

Introduction

Japanese compound surnames exhibit complex phonological patterns with respect to prosody and the application of a voicing alternation known as *rendaku*. This dissertation gives a full account of the sound patterns of Japanese surnames, which holds various implications for phonological theory. This chapter introduces the main issues to be addressed in the study and gives a preview of the analysis to be proposed. I will also give a sketch of Japanese phonology, illustrating the transcription system used in the study.

1.1 Issues

1.1.1 Peculiar rendaku and accent patterns in surnames

Japanese surnames are mostly compounds. Just like regular compound words, they may undergo rendaku, a phonological alternation which voices the initial consonant of the second element of a compound (e.g. /yama/ + /ta/ → [yama-**d**a] 'mountain-paddy'; see Section 2.2 for more details). However, it has been observed in the literature that rendaku application in surnames is somewhat different from that in regular compounds (see Sugito 1965; Kubozono 2005; Zamma 2005; Asai 2014 among others). For example, the presence of a voiced obstruent in the first element inhibits the voicing alternation; such a pattern is not attested in rendaku in regular compounds (Irwin 2014a,b; Sano 2015). (Also see Section 2.3 for the details and other peculiarities in rendaku in surnames.) Surnames have also been described as showing unusual prosodic patterns. Though they are compounds, they do not follow the usual compound accent rule, and the location of accent in this word class is largely predictable (see e.g. Ito and Mester 2016). Another prosodic charac-

teristic seen in compound surnames is an inverse correlation between accentedness and rendaku application: names with rendaku voicing tend to be unaccented while those without voicing are often accented. No such correlation is found in regular compounds (see Alderete 2015). (See Section 2.4 for the details of the prosodic characteristics of surnames.) Despite all these descriptions of how the phonological patterns of compound surnames differ from those of regular compounds, no study has ever provided an explanation as to *why* surnames show those peculiarities. The first goal of the dissertation is thus to propose a principled account of the rendaku and accent behaviors of compound surnames.

1.1.2 The coexistence of lexicalization and productivity

Rendaku in surnames poses a challenge for current theories of phonology. As will be shown below, the phenomenon is very much affected by both general phonological factors and lexical factors. The results of an experiment using nonce surnames as the stimuli suggest that rendaku application is a productive process (Section 4.3). Japanese speakers apply voicing to a given surname based on its phonological configuration. Variable application patterns also suggest that various factors are in play and interact with each other. The role of those phonological factors turn out to be more important in rendaku in surnames than in rendaku in regular compound words. However, a close look at the patterns of existing surnames suggest that rendaku is also highly conditioned by the idiosyncratic properties of each surname. Given two surnames which are phonologically similar but have different profiles for rendaku realization (e.g. /saka-ta/ [saka-ta] 'slope-paddy' and /taka-ta/ [taka-**d**a] 'high-paddy'), it is virtually impossible to explain why one shows voicing and the other does not; it is simply determined so by their lexical properties. A full analysis of the voicing alternation, however, must be able to predict such fixed rendaku patterns of existing surnames (or rendaku patterns with some fixed rates for those showing variation: e.g. /naka-ta/ → [naka-ta]: 80.4%, [naka-da]: 20.6%). (See Section 5.2 for discussion.) It is this coexistence of phonologically-driven productivity and lexical conditioning which makes the modeling of rendaku in surnames particularly challenging. The second goal of the dissertation is to model both aspects of the phenomenon.

1.2 Preview of the proposals

1.2.1 Representation: A stem composed of stems

In order to account for the peculiarity of their rendaku patterns, I propose that compound surnames are represented in the grammar as single stems which are in turn composed of stems. This recursive morphological structure of a compound surname differs from that of a regular compound word, as shown in (1).

(1) Representations of compound surnames and regular compounds

a. A compound surname: A stem composed of stems
$[\ A\ ]_{stem} \quad + \quad [\ B\ ]_{stem} \quad \rightarrow \quad [\ [\ A\ ]_{stem}\text{-}[\ B\ ]_{stem}\ ]_{stem}$

b. A regular compound: Two combined stems
$[\ A\ ]_{stem} \quad + \quad [\ B\ ]_{stem} \quad \rightarrow \quad [\ A\ ]_{stem}\text{-}[\ B\ ]_{stem}$

Base on the representation, I further claim that the rendaku and prosodic patterns of compound surnames are driven by stem-internal phonology. The proposal will give a unified account of their peculiar sound patterns. For example, the fact that voiced obstruents in the first element block the voicing alternation can be explained by a constraint which bans multiple voiced obstruents within a stem. Since a compound surname is represented as a stem, the stem-internal restriction applies to the whole compound, blocking rendaku voicing which would otherwise create multiple voiced obstruents *stem-internally*. Note that the ban on the cooccurrence of voiced obstruents in a stem is motivated independently of rendaku in surnames and its effects are found elsewhere in the language (see e.g. Ito and Mester 1986). It will be shown that most of the phonological patterns which are unique to compound surnames can be attributed to stem-internal phonology. (See Chapter 3.) The validity of the claim will also be corroborated by the results of a corpus study and an experiment. (See Chapter 4)

1.2.2 Grammar: A Maximum Entropy model with lexically-specific constraints

In order to meet the challenge of capturing both the lexicalized aspect and the productive aspect of rendaku application in surnames, I propose a Maximum Entropy Harmonic Grammar model composed of general phonological constraints and what we call "lexically-specific constraints" (Moore-Cantwell and Pater 2016). A Maximum Entropy Harmonic Grammar (MaxEnt HG) refers to a log-linear model couched within the framework of Harmonic Grammar (Legendre et al. 1990), which employs Optimality-Theoretic constraints (Prince and Smolensky 1993/2004) with constraint weights rather than rankings. (See Section 5.3.1 for more detailed descriptions of the framework.) When provided with a set of constraints, a MaxEnt HG model can learn and generate a probability distribution over a set of phonological forms based on their constraint violation profiles. The model's basic architecture allows for capturing so-called token variation; it can assign probabilities to output candidates of a given input form (e.g. /naka-ta/ → [naka-ta]: 80.4%, [naka-da]: 20.6%). Furthermore, when the model is equipped with constraints indexed to every lexical item, it can learn the idiosyncratic behavior of individual items. This will help account for the lexically-conditioned patterns of rendaku in surnames. I will also impose learning biases which work against these lexically-specific constraints to ensure that the model still prefers to learn general phonological patterns in the data. This is crucial for predicting the rendaku patterns of nonce surnames. Through learning simulations, I will show that a MaxEnt model with general constraints, lexically-specific constraints and learning biases can not only accurately learn the lexically-conditioned rendaku patterns of existing surnames but also capture the general patterns of productive rendaku application in non-existing surnames. (See Section 5.4.)

1.3 Notation and a sketch of Japanese phonology

1.3.1 Notes on transcriptions

For transcriptions, I will use a modified version of the Kunrei romanization system, which is commonly used in the literature on Japanese phonology. There are several differences between the

adopted system and the International Phonetic Alphabet. I will note some of these differences when I introduce phonemes below. Both the phonemic forms and the phonetic forms of surnames presented as data will be transcribed using the Kunrei-based system (e.g. /taka-sima/ [taka-sima] 'high-island'). Thus, the transcriptions in this study are generally very broad; this is meant to reduce phonetic details which could distract the reader from relevant discussion. When relevant, Chinese characters, or *kanji*, will also be supplied (e.g. /taka-sima/ [taka-sima] 高島 'high-island'). When the name of an individual is given in the text, it will be italicized and written using the Hepburn romanization system, which is largely based on English orthography and is conventionally used in the transliteration of Japanese names (e.g. *Takashima* 'high-island'). The names of Japanese scholars in citations are also usually in Hepburn romanization but shown without italicization (e.g. Takashima 2017).

1.3.2 Segments: Phoneme inventory

The main phonemes of the Japanese language are given in (2) below.

(2) Main phonemes of Japanese

Vowels		Consonants							
i	u	p	b	t	d		k	g	
e	o			s	z				h
a			m		n				
			w		r	y			

Vowel length is phonemic. A long vowel is indicated by doubling the symbol as in [aa]; the IPA equivalent would be [aː]. A phonetic realization of /u/ is often quite centralized and unrounded; in IPA, the sound is commonly transcribed as [ɯ].

The palatal approximant is transcribed as /y/ in this study; its IPA equivalent would be /j/. /r/ is often realized as a tap or a (retroflexed) flap. Most of the consonants have their palatalized counterparts (with some restrictions of vowel contexts). Palatalized consonants are usually considered to be independent phonemes, but they are written with two symbols in this study: e.g. /ky/ and

/ny/. Most consonants can also be geminated. A geminate consonant is indicated by doubling the symbol as in [kk] and [nn]. Some analyses posit /N/ as an independent phoneme. However, the distribution of the sound is limited to syllable-final position and it may be analyzed as any of the nasal sound underlylingly. In this study, the alveolar /n/ will be used for both the phonemic and phonetic forms of this moraic nasal sound: e.g. /san/ [san] 'three.'[1]

The alveolar fricatives /s/ and /z/ become alveolo-palatal before the high front vowel /i/, which is often described under the label "palatalization." Note that such palatalization will not be reflected in the transcriptions here, as in /si/ [si] and /zi/ [zi]; the IPA equivalents of the phonetic forms would be [ɕi] and [ʑi]~[dʑi] respectively.[2] /t/ and /d/ are also palatalized involving affrication before /i/. Again, these allophonic alternations will not be reflected in the transciptions, as in /ti/ [ti] and /di/ [di]; more narrowly, the phonetic forms would be [tɕi] and [dʑi]~[ʑi] respectively. /t/ and /d/ also become affricates before /u/; this is also not transcribed, as in /tu/ [tu] and /du/ [du], although in IPA they would be [tsɯ] and [dzɯ]. Phonemically palatalized coronal obstruents, transcribed as /sya/ [sya] and /tya/ [tya], are also realized as alveolo-palatal (with affrication for the stops /ty/ and /dy/); narrower transcriptions of the examples would be [ɕa] and [tɕa] respectively. /h/ becomes palatal before /i/ or when phonemically palatalized and bilabial before [u]. They are transcribed as [hi], [hya] and [hu] in this study; the IPA equivalents would be [çi], [ça] and [ɸɯ] respectively.

1.3.3 Prosody: Pitch accent

Japanese has lexical pitch accent. Words can be contrasted by the presence of accent or the location of accent. Phonetically, accent is realized as a fall in pitch, which can be represented as a high-low contour, associated to the accented syllable. In transcription, an accented syllable is indicated

[1]The voiceless bilabial fricative (/ɸ/ in IPA) may also be claimed to be an indendent phoneme, given that it may appear in any vowel context in loanwords. I omit it from the table of the main phonemes, since it only appears allophonically before /u/ in the native stratum and the focus of this dissertation is the phonology of native Japanese words.

[2]/zi/ and /di/ can each be variably realized as [ʑi] or [dʑi] (IPA), neutralizing the contrast between /z/ and /d/ in the context.

by an accute accent mark, as in [á]. For clarity, I also adopt Ito and Mester's (2016) notation of using a superscript number to indicate the location/presence of an accented "mora," with moras being counted from the end of the word. For example, a superscript '2' indicates that accent is on the penultimate (second-to-last) mora, as in ^{2}[hási] 'chopsticks,' and a superscript '1' indicates that accent is on the last mora, as in ^{1}[hasí] 'bridge.' A superscript '0' indicates that the word is unaccented (i.e. no accented mora), as in ^{0}[hasi] 'edge.' Pitch accent is not marked when it is not under discussion.

CHAPTER 2

Data: Rendaku and accent in Japanese surnames

2.1 Chapter overview

This chapter presents issues surrounding the rendaku and accent patterns of compound surnames in detail. Section 2.2 gives a description of the phenomenon of rendaku application and its blocking by Lyman's Law in compounds in general. In Section 2.3, I present segmental factors affecting rendaku in compound surnames, pointing out the peculiarity of the voicing alternation in proper nouns as compared to that in common nouns. Section 2.4 discusses the prosodic patterns of surnames in association with rendaku. Section 2.5 describes how rendaku application is affected by lexical factors. Section 2.6 presents the history of surnames as well as rendaku voicing in Japanese and makes a claim that rendaku in surnames should be understood not as a vestige of old sound patterns but as a productive phonological process in the synchronic grammar of Japanese.

2.2 Background

2.2.1 Rendaku and lexical propensities

Rendaku, or sequential voicing (Martin 1952), is a morphophonological process in Japanese whereby the initial obstruent of the second member of a compound becomes voiced, as shown in (3) below.[1]

[1]To be more precise, rendaku may occur not only in noun-noun compounds but also in morphologically-complex words of which one of the members is an affix or affix-like element (see Irwin 2012; Vance 2015). In this dissertation, I use the term "compound" for simplicity. Also, if a word is composed of more than two elements, not only the second element, as stated in the definition here, but also other non-initial elements may undergo rendaku, arguably under some

(3) Compound formation involving rendaku

a.	maki	+	s̲usi	→	maki-**z**usi	‘rolled-sushi’
b.	hon	+	t̲ana	→	hon-**d**ana	‘book-shelf’
c.	ori	+	k̲ami	→	ori-**g**ami	‘folding-paper’
d.	osi	+	h̲ana	→	osi-**b**ana	‘pressed-flower’

As can be seen, if the second element of a compound (henceforth referred to as E2) starts with a voiceless obstruent such as [s], [t], [k] or [h], the consonant becomes voiced as a result of compounding. Note that [h] alternates with [b] for historical reasons.[2] Rendaku applies vacuously when E2 starts with a sonorant (e.g. maki + m̲ono → maki-**m**ono ‘rolled-material’).

Despite being one of the most well-studied phenomena of sound alternations in Japanese phonology, rendaku remains a recalcitrant problem, mainly due to irregularities in its application. McCawley (1968), one of the oldest studies on Japanese phonology in generative linguistics, writes “I am unable to state the environment in which the ‘voicing rule’ applies [...]” (McCawley 1968:87). Some studies have shown that its applicability is affected by factors such as lexical strata, the length of the components and the morphological or semantic structure of the compound (see Vance 2015 among others for a recent comprehensive overview). Other studies have also pointed out a morpheme-specific nature of rendaku. Rosen (2001) argues that some morphemes, when they appear as E2 of a compound, often undergo the voicing alternation (hence he calls them “rendaku lovers”) while some others typically do not (“rendaku haters”) or even never (“rendaku immune morphemes”). (Also see Nakagawa 1966; Martin 1987; Vance 1979, 1980b for similar observations.) In other words, all else being equal, whether rendaku applies or not is determined by each E2’s lexical propensity to undergo the voicing alternation.

The lexical propensities of E2 morphemes still do not give a full account of the phenomenon, however. Rosen (2001) classifies morphemes as “rendaku lovers” if they voice in more than 66%

additional conditions (see Otsu 1980 for the so-called Right Branch Condition; see Vance 1980a for criticisms; also see (Kawahara 2015a) and references therein for discussion).

[2]It is generally agreed that the /h/ in present-day Japanese was derived from Proto-Japanese */p/ through /ɸ/ (Ueda 1898 via Takayama 2012; also see Kiyose 1985; Miyake 2003:66–77,164–166.)

of the compounds they occur in, and as "rendaku haters" if they voice in less than 33% of the compounds they occur in.[3] Irwin (2016a) further adds two other categories: "rendaku submissive" morphemes which always undergo rendaku, and "rendaku waverers" which fall between "lovers" and "haters," voicing from 34% to 66% of the time. Notice that, although such classification may help capture the tendencies of rendaku application, it cannot be used as a crucial factor in determining whether or not a given compound undergoes rendaku. There are cases where the exact same E2, whether a lover, waverer or hater, undergoes rendaku in one compound but not in another for no specific reason. This irregularity of rendaku is exemplified in (4) by compounds with /kuti/ 'mouth, entrance' as E2. /kuti/ is a rendaku lover according to Rosen (2001), but it still shows some variability in rendaku application. To the best of my knowledge, none of the factors that have been proposed in the literature can truly explain why rendaku applies in (4a), but not in (4b), and why (4c) shows free variation (some examples are taken from Vance 2005c).[4]

(4) Irregularity of rendaku

a.	kage	+	kuti	→	kage-**g**uti	'behind-mouth (gossip)'
b.	haya	+	kuti	→	haya-**k**uti	'fast-mouth (fast speech)'
c.	waru	+	kuti	→	waru-**k**uti ~ waru-**g**uti	'bad-mouth (slander)'

This suggests that rendaku application may be determined on a word by word basis. The lexical irregularities of rendaku have led some linguists to conclude that compound voicing is not a productive rule in the phonological grammar of Japanese and that the rendaku patterns of all compounds are simply lexicalized. A strong argument against this view is that rendaku is attested in newly coined words and also in experiments using nonce words (Vance 1979, 1980b; Ihara and Murata 2006; Kawahara and Sano 2014a,b,c to name a few; also see Kawahara 2016 and references

[3]Rosen limits his data to two-element compounds where both elements are bimoraic native Japanese morphemes, when he calculates the rendaku application rate of each morpheme. He claims that when one of the elements exceeds two moras, compounds are generally more likely to undergo rendaku. See Section 2.4.3 for more details.

[4]Sato (1989) notes that rendaku occurs less often if the first element (E1) is an adjectival element. This could possibly explain why (4a) does not undergo rendaku. Note, however, that (4c) also has an adjectival root as E1 and it still variably undergoes rendaku. Sato admits that this is just a tendency and there are actually many counter-examples to this generalization.

therein). This being said, there is no doubt that rendaku has many irregularities as stated in Vance (2014) and its status in phonology has always been debated. The reader is referred to Fukuda and Fukuda (1994), Kawahara (2015a, 2016), Kubozono (2005), Ohno (2000) and Vance (2014, 2015) among others for relevant discussion.

2.2.2 Lyman's Law and its role in the lexicon

Although the phonological status of rendaku is highly controversial, there is one point about its applicability that has reached a close consensus among linguists; the voicing alternation does not occur if E2 already contains a voiced obstruent. This blocking of rendaku is commonly known as Lyman's Law (Lyman 1885, 1894).[5] See the examples in (5) below.

(5) Lyman's Law blocks rendaku

	a.	yama	+	k͟az͟e	→	yama-**k**az͟e,	*yama-**g**az͟e	'mountain-wind'
cf.	b.	yama	+	k͟umo	→	yama-**g**umo		'mountain-cloud'
	c.	oo	+	t̲okag̲e	→	oo-**t**okag̲e,	*oo-**d**okag̲e	'big-lizard'
cf.	d.	oo	+	t̲anuki	→	oo-**d**anuki		'big-racoon'

In (5a), rendaku does not apply since the underlying form of the E2 /kaze/ 'wind' has /z/ in the second syllable. This contrasts with (5b) where the E2 /kumo/ 'cloud' does not contain a voiced obstruent, and the compound does undergo rendaku. Note that this blocking of voicing by a voiced obstruent is not a local phenomenon. As shown in (5c), the /g/ in /tokage/ 'lizard' prevents voicing of the initial /t/ although they are separated by another syllable. The law can be viewed as a kind of Obligatory Contour Principle effect on the stem level (Ito and Mester 1986); it disallows the occurrence of multiple voiced obstruents in the domain of a single stem. Rendaku is

[5] According to Miyake (1932), Motoori Norinaga (1730–1801), a Japanese philologist and physician in the Edo period, also documents this blocking factor of rendaku in his *Kojiki-den* (Motoori 1790-1822) (see Miyake 1932; Endo 1980; Vance 1980b). Suzuki (2004) also points out that Kamo no Mabuchi (1697–1769) has some notes on the phenomenon in his *Goi-kō* (Kamo 1765-1789). Kamo mentions that compounds like /yama-kaze/ do not undergo rendaku as the second element contains a voiced obstruent. Although the law could be called the "Kamo-Motoori-Lyman Law" to acknowledge the three scholars who may have discovered it independently, I will refer to it as Lyman's Law in this dissertation following the convention in the generative linguistics literature.

thus blocked when it would create a second voiced obstruent in E2.

Lyman's Law is essentially inviolable, allowing only a handful of exceptions.[6] There is also evidence for its psychological reality. Although the effect is often gradient in experimental settings, Japanese speakers show dispreferences for nonce compound words with rendaku violating Lyman's Law (see Vance 1979, 1980b; Ihara et al. 2009; Kawahara 2012; Kawahara and Sano 2014a,b, 2016 among others).[7]

The statement that there should not be multiple voiced obstruents within a stem also holds true as a static generalization of the Japanese lexicon. Native Japanese stems do not usually contain more than two voiced obstruents except those that are onomatopoeic or those with pejorative connonations. Lyman's Law is thus not just a rendaku blocking factor but a more general phonotactic constraint on native stems in Japanese (see Morita 1977; Ito and Mester 1986; Yamaguchi 1988).[8]

2.2.3 Compound surnames and rendaku

Most Japanese surnames are compounds. Of the top 10,019 names in a large database of surnames (Shirooka and Murayama 2011), 96.38% of them are composed of multiple morphemes: 3.62% are monomorphemic (e.g. *Hara* /hara/ 原 'field'), 88.75% are bimorphemic (e.g. *Tahara* /ta-hara/ 田原 'paddy-field'), 7.58% are trimorphemic (e.g. *Odawara* /o-ta-hara/ 小田原 'small-paddy-field')[9] and 0.05% are composed of four morphemes (e.g. *Teshigawara* /te-si-kawa-hara/ 勅使河原

[6]See Kindaichi (1976/2005:342), Vance (1987:137) and Vance (2014:159) for the few exceptional cases. Also see Suzuki (2005, 2008) for a historical explanation on how such words arose.

[7]Note, however, that Vance (1979, 1980b) finds variation among participants in how sensitive they are to Lyman's Law in the wug test he conducted. He concludes that the law is psychologically real only for some speakers of present-day Japanese.

[8]The ban on multiple voiced obstruents within a stem does not have such categorical effects in loanwords as it is violated by a number of loan stems (e.g. [b̲ag̲u] 'bug'; Ito and Mester 2003). However, the effects of Lyman's Law may show up even in loans under certain conditions. See Nishimura (2003); Kawahara (2006, 2011, 2012) among others for the details.

[9]The alternation of /ta/ to [**d**a] is due to rendaku application. /hara/ to [wara] is a case of vocalization of [h], which happens occasionally in surnames. See Akinaga (1977b); Asai (2014); Vance and Asai (2016) for details.

'imperial-envoy-river-field').[10] In writing, each morpheme is represented by a Chinese character or *kanji*. For example, the fourth most common surname *Tanaka* /ta-naka/ 田中, which literally means 'paddy-center,' is written with two Chinese characters and can be decomposed into /ta/ 田 'paddy, rice field'[11] and /naka/ 中 'center, middle, inside.'

Compound surnames qualify for undergoing rendaku if the second element (E2) starts with a voiceless obstruent. As in regular compounds, rendaku in surnames is not an iron-clad rule. Certain names typically show the voicing change while others do not, and there are also some that show "free variation."[12] This is illustrated in (6) below with surnames which have /ta/ 田 'paddy' as E2.

(6) Japanese surnames and rendaku

a.	yama	+	t̲a	→	yama-**d**a	山田	'mountain-paddy'
b.	nari	+	t̲a	→	nari-**t**a	成田	'completed-paddy'
c.	naka	+	t̲a	→	naka-**t**a ~ naka-**d**a	中田	'center-paddy'

It is worth noting here that rendaku is usually not reflected in the orthography. On most occasions, surnames are written in the kanji script, where each character represents one morpheme possibly with multiple different readings. For example, in (6), the second element is always written with the same character 田 'paddy,' whether it is realized as [ta] or [da].

[10]The morphological structures of surnames are inferred from the orthography on the assumption that each morpheme is represented by one Chinese character. The assumption is not necessarily true, since there are a handful of morphemes which can be written with either one character or two characters (e.g. /kubo/ '(geologic) depression': 窪 or 久保). The meanings of the example surnames are literal translations.

[11]The morpheme originally means 'field' more generally, and 'paddy, rice field' may not be an accurate translation to be used for certain surnames with /ta/. However, I will use 'paddy, rice field' in order to differentiate it from other morphemes also meaning 'field,' such as /hara/ '(natural) field' and /hata/ '(plowed) field.'

[12]"Free variation" here does not mean that one single person's surname can be pronounced either with or without rendaku at each utterance. It rather means that we find both people who have the rendaku reading and people who have the non-rendaku reading even though their surnames are composed of the same morphemes and written with the same kanji. I will use the term "free variation" for the lack of better words.

2.2.4 Summary

To summarize, rendaku is a morphophonological process that voices the initial segment of the second element (E2) of a compound. Although its application is conditioned by a number of factors including lexical propensities and is often said to be unpredictable, it is categorically blocked by Lyman's Law, which prohibits the occurrence of multiple voiced obstruents within a stem. Most Japanese surnames are compounds and thus may undergo rendaku like regular compound words.

2.3 Segmental factors

This section describes how segmental factors affect rendaku application in compound surnames. I will show that the rendaku patterns in surnames are peculiar and different from those in regular compounds.

2.3.1 Strong Lyman's Law?

Although the rendaku patterns in compound surnames may appear to be as complex as those in normal compounds, Sugito (1965) reveals that the voicing alternation is largely predictable when it comes to surnames with /ta/ as E2. She observes that rendaku is conditioned by the onset consonant of the syllable preceding /ta/, namely the last consonant of the first element (henceforth E1).[13] For example, if E1's last consonant is /s/, /t/, /k/, /m/ or /n/, rendaku is commonly observed, as in (7a). On the other hand, if the consonant is a voiced obstruent such as /z/, /b/, /d/ or /g/, rendaku never applies, as in (7b).[14]

[13]Given that Japanese allows only a limited set of coda consonants (see Ito 1986; Ito and Mester 1994), "E1's last consonant" usually refers to "the onset consonant of E1's final syllable." Note, however, that the final syllable of E1 may have a moraic nasal or the first half of a geminate consonant in coda (e.g. kan-**d**a, nit-**t**a), in which case I use the term "E1's last consonant" to refer to the coda consonant instead of the onset consonant.

[14]Sugito's (1965) generalizations are not based on features and she also includes the glide /y/ in the set of consonants that block rendaku as in (7b). Here, I follow the generalizations made by Kubozono (2005) and Zamma (2005), who analyze Sugito's (1965) data.

(7) E1's last consonant conditions rendaku in surnames with /ta/

a. If /s, t, k, m, n/, mostly rendaku:
   a̲sa-**d**a ma̲tu-**d**a ta̲ka-**d**a ya̲ma-**d**a sa̲na-**d**a

b. If /b, d, g, z/, no rendaku:
   ka̲zi-**t**a si̲ba-**t**a ka̲do-**t**a naga̲-**t**a

The table in (8) gives the number of surnames with /ta/ as E2 showing either rendaku or no rendaku (i.e. [**d**a] or [**t**a]) sorted by the type of the last consonant in E1 in the data compiled by Sugito (1965). As can be seen, rendaku applies in most surnames which have a voiceless obstruent or a nasal in E1's final syllable, while there is no surname with a voiced obstruent in E1 which shows rendaku voicing. (Note that Sugito excludes surnames showing free variation as in (6c) from her data. Also, her survey is only concerned with surnames with an E1 morpheme that is bimoraic.[15])

(8) No. of surnames with /ta/ by E1's last consonant and rendaku (Sugito 1965)

E1 last cons.	**t**a	**d**a
/s, t, k, m, n/	16	158
/z, b, d, g/	30	0

This blocking of rendaku by a voiced obstruent in E1 may, at first glance, seem to be the application of Lyman's Law; one may think that if E1 already contains /z/, /b/, /d/ or /g/, rendaku should be blocked as it would create two voiced obstruents within the compound (e.g. si̲ba + t̲a → *si̲ba-**d**a, si̲ba-**t**a). Note, however, that the law's effect is normally stem-bounded. Rendaku is blocked if there is /b/, /d/, /g/ or /z/ in *E2* since voicing its initial consonant would create two voiced obstruents within that stem. Voiced obstruents in *E1*, on the other hand, should not affect rendaku application across a morpheme boundary. Examples of regular compounds are shown in (9).

[15]If E1 is longer than two moras, rendaku is generally more likely to apply regardless of the voicing of E1's last consonant (Zamma 2005); e.g. /yanagi-ta/ [yanagi-da] ~ [yanagi-ta]. See Sections 2.4.3 for more details.

(9) No rendaku blocking by voiced obstruents in E1 in regular compounds

	a.	a̲ge	+	s̲oko	→	a̲ge-**z**oko		‘raised-bottom’
	b.	ta̲bi	+	h̲ito	→	ta̲bi-**b**ito		‘travel-person’
	c.	ka̲be	+	k̲ami	→	ka̲be-**g**ami		‘wall-paper’
cf.	d.	kuro	+	k̲ab̲i	→	kuro-**k**ab̲i,	*kuro-**g**ab̲i	‘black-mold’

It has been proposed in the literature that, even though examples like (9a), (9b) and (9c) are found, rendaku is still less likely to occur when there is a voiced obstruent in E1 (see Kindaichi et al. 1988:264; Sato 1989; also see Takayama 1992; Ito and Mester 2003:110; Labrune 2012). Although the proposal has never been formally stated, one logical interpretation of the idea is that Lyman’s Law, which is usually stem-bounded, also shows some gradient effects on the word level. This extended version of Lyman’s Law is often referred to as “Strong Lyman’s Law.”[16] It is known that phonotactic restrictions that have a categorical effect within a smaller domain (e.g. stem) also have a weaker effect in, or “leak to,” a bigger domain (e.g. word) in some languages (Martin 2007, 2011). Strong Lyman’s Law may then be another case of leakage of stem-internal phonotactics.

This hypothesis, however, does not seem to be borne out. According to recent studies, in regular compounds (excluding compounds of proper nouns such as surnames), voiced obstruents in E1 do not affect the likelihood of rendaku application. Using a large corpus of Japanese compounds (Irwin and Miyashita 2013-2016; Irwin et al. 2017), Irwin (2014a) shows that there is no statistically significant tendency that rendaku applies less when E1 contains a voiced obstruent. Irwin (2016a) corroborates the finding by conducting more detailed analyses of the same data; the two voiced obstruents /b/ and /z/ in E1 do make the rendaku application rates lower than expected when compounds with high word frequency are concerned, but voiced obstruents overall still do not affect rendaku considerably. Sano (2015) reaches the same conclusion based on a study using a different corpus (the National Institute for Japanese Language 2012).[17] In contrast, Asai (2014)

[16]The term “strong version of Lyman’s Law” was originally used by Unger (1977) to refer to the restriction on the occurrence of multiple voiced obstruents within a word in Old Japanese. See Section 2.6.1 for the details.

[17]In fact, when Lyman (1894) first described what would be later known as Lyman’s Law, he also made it clear that a voiced obstruent in the syllable preceding E2 (i.e. E1’s final syllable) does not affect rendaku based on his own observation of words in *Waei Gorin Shūsei*, the oldest Japanese-English dictionary (Hepburn 1867/1872; second

finds some weak effects of Strong Lyman's Law using data compiled from magazines. We should note, however, that Asai's (2014) corpus is smaller than the ones used in Irwin (2014a, 2016a) and Sano (2015), and it also includes person names and place names, which could well be confounding factors. (See Section 2.3.3 for relevant discussion.) The psychological reality of the extended version of Lyman's Law is also questionable. Kawahara and Sano (2014c) find no clear experimental evidence that such patterns are internalized in the grammar of Japanese speakers.

The question now arises as to why compound surnames are subject to Strong Lyman's Law, which has at best minimal effects in normal compounds. Sugito's (1965) data clearly show that a voiced obstruent in E1 blocks rendaku application in surnames with /ta/ as E2. Subsequent studies (Asai 2014, Kubozono 2005 and Zamma 2005 among others) further report that Strong Lyman's Law plays a more general role in rendaku in surnames; although its effects may not be as strong as in the case of surnames with /ta/, a voiced obstruent in E1 makes rendaku application less likely in surnames with other E2s such as /sima/ and /kawa/.[18] If both compound surnames and normal compounds may undergo rendaku, why is the domain of Lyman's Law different (that is, the word in the former but the stem in the latter)? What mechanism derives such difference? As will be shown below, the difference lies in the way the two types of compounds are treated in morphophonology. I propose in Section 3.2.2 that compound surnames composed of stems are recursively treated as "stems" in the grammar. As a consequence, the ban on multiple voiced obstruents applies to the whole compound in the case of a surname, deriving effects of Strong Lyman's Law.

2.3.2 Peculiarity of /k/ in the first element

Rendaku in surnames shows further complications, one of which is that, among voiceless obstruents, /k/'s in E1 uniquely allow non-application of rendaku in surnames with /ta/ as E2 (Kubozono

edition): "A sonant [(voiced sound)] in the syllable before has no effect on the nigori [(voicing)] (about 150 words with, and about 150 without)." (Lyman 1894:162; explanation of some terms was added in square brackets).

[18]Kubozono (2005) and Zamma (2005) also note that certain E2 morphemes such as /kuti/ are not sensitive to Strong Lyman's Law and undergo rendaku regardless of the voicing of E1's last consonant. See Section 2.5 below for E2's lexical propensities in compound surnames.

2005; Zamma 2005). As shown by the examples in (10a), rendaku is more common in surnames with /k/ in E1; however, there are also quite a few names showing no rendaku, as in (10b), or those showing variation, as in (10c).

(10) Rendaku in /ta/-surnames with /k/ in E1's final syllable

a.	hu<u>k</u>a-**d**a	o<u>k</u>a-**d**a	ta<u>k</u>e-**d**a	i<u>k</u>e-**d**a	ta<u>k</u>a-**d**a	to<u>k</u>u-**d**a
b.	sa<u>k</u>a-**t**a	o<u>k</u>i-**t**a	a<u>k</u>i-**t**a	i<u>k</u>u-**t**a	ta<u>k</u>i-**t**a	
c.	na<u>k</u>a-**t**a ~ na<u>k</u>a-**d**a		hu<u>k</u>u-**d**a ~ hu<u>k</u>u-**t**a			

This contrasts with surnames having /s/ or /t/ in E1. The table in (11) shows the number of /ta/-surnames with and without rendaku ([**t**a] or [**d**a]) sorted by the place of a voiceless obstruent in E1's final syllable (/s/, /t/ or /k/) in Sugito's (1965) data.[19] As can be seen, /s/ and /t/ always trigger rendaku while /k/ allows both rendaku and non-rendaku forms. (Note that variable names like the ones in (11c) are not included in the table here, as mentioned above.)

(11) No. of /ta/-surnames by E1-voiceless obstruent and rendaku (Sugito 1965)

E1 cons.	**t**a	**d**a
/s/	0	35
/t/	0	26
/k/	13	31

Both Kubozono (2005) and Zamma (2005) note the oddity of the data pattern. Kubozono states "The reason for this peculiar behavior of /k/ remains unclear." (Kubozono 2005:10). Zamma also writes "Exceptions to [Sugito's (1965)] generalization seem rather abundant, but if we assume /k/ can be exceptionally regarded as voiced — as does Kubozono (this volume) — [most of the data points] can be correctly predicted [...]" (Zamma 2005:160).

Particularly relevant to us here is the fact that this peculiar behavior of /k/ in E1's final syllable is not seen in the rendaku patterns in regular compounds. Corpus studies by Irwin (2014b, 2016a) show that no particular consonant in E1's final syllable exerts any exceptional effect on the rate of

[19] I follow the format of the table in Kubozono (2005), who reports Sugito's (1965) data with some slight modification in the presentation.

rendaku application in normal compounds.[20] Why, then, does the place of a voiceless obstruent in E1, particularly /k/, affect rendaku in compound surnames? It will be shown below that the peculiarity of /k/ also follows from the principle that compound surnames are treated as single stems. I will argue in Section 3.3.2 that surnames are subject to the phonology of stems despite their structure as compounds, and that /k/'s behavior in surnames with E2-/ta/ is simply a reflection of Identity Avoidance, which is one of the major stem-internal phonotactic restrictions in Japanese.

2.3.3 Other patterns: Sonorants in E1

Another peculiarity found in rendaku in surnames is that sonorants in E1 also affect the likelihood of voicing. Kubozono (2005) observes that, in Sugito's (1965) data of surnames with /ta/ as E2, the approximants /r/, /y/ and /w/ in E1 tend to block rendaku.[21] Other studies also claim that /r/ in E1 specifically makes the voicing alternation less likely in surnames (Hirata 2010; Asai 2014). As shown in (12), most of the common surnames with /r/ in E1's final syllable and /ta/ as E2 resist rendaku, with just a few exceptions, as in (12b).

(12) /r/ in E1 tends to block rendaku

a. nar̲i-**t**a kur̲i-**t**a kur̲a-**t**a mur̲a-**t**a mor̲i-**t**a hir̲o-**t**a ar̲i-**t**a ...
b. har̲a-**d**a kur̲o-**d**a ter̲a-**d**a

Given in (13) is a table showing the number of /ta/-surnames with and without rendaku sorted by the type of E1-approximant in Sugito's (1965) data.[22] As can be seen, all approximants, not just /r/, seem to inhibit rendaku when they appear in E1's final syllable.

[20] Asai's (2014) corpus study, on the other hand, shows that the place of the consonant in E1's final syllable affects rendaku application. Again, note that Asai (2014) includes not only common nouns but also proper nouns in his data. This makes the results very difficult to interpret, as we are particularly interested in the comparison between the two types of compounds.

[21] It may be more accurate to refer to the three sounds as "non-nasal sonorants" (Ito and Mester 2003), given that /r/ is most typically realized as a tap or flap, as noted in Section 1.3.2. In this dissertation, I use the term "approximants" for simplicity.

[22] The table is adapted from Kubozono (2005) who reports Sugito's (1965) data with some modifications in the presentation. Note that Sugito excludes names showing free variation such as [toy̲o-**t**a] ~ [toy̲o-**d**a].

(13) No. of /ta/-surnames by E1-approximant and rendaku (Sugito 1965)

E1 cons.	**t**a	**d**a
/r/	31	3
/w/	7	2
/y/	8	0

There has been a claim that /r/ in the final syllable of E1 dampens rendaku also in regular compounds (Hirano 2013, via Irwin 2016a:97; also see the discussion in Vance and Asai 2016:129–130). Irwin (2016a) calls it into question, noting that the rendaku rates are only slightly lower than expected when E1's last consonant is /r/ in his large corpus data. Toda (1988:89) investigates the rendaku patterns of regular compounds in Late Middle Japanese (12c.–16c.) and Early Modern Japanese (17c.–19c.) (see Section 2.6.1 for more details of her corpus studies). She states that, although /r/ does make the rendaku rate relatively low, it does not inhibit the voicing alternation considerably. On the other hand, Asai (2014) finds that /r/ in E1 does inhibit rendaku significantly in his magazine-based corpus of present-day Japanese compounds (though it is far from being a strong blocking factor). The results should be taken with a grain of salt, however, if we are to attribute these rendaku patterns to common nouns. Since Asai's (2014) database includes proper nouns such as person names and place names, the effect he finds may come from these compounds. That is, the results are still compatible with the other studies showing that /r/ in E1 has at best minimal effects on rendaku application in regular compounds (Toda 1988; Irwin 2016a).[23] These rendaku patterns of regular compounds with respect to E1-approximants are in contrast to what we see in surnames. On the assumption that the patterns in Sugito's (1965) data of surnames with E2-/ta/ presented in (13) apply to surnames in general, why do approximants in E1 block rendaku in compound surnames, but not in regular compounds?

Turning to the other type of sonorants, surnames with nasals in E1 commonly show rendaku. Recall that Sugito (1965) classifies E1-nasals as consonants that trigger rendaku along with voice-

[23]In addition, /y/ and /w/ do not seem to have remarkable dampening effects on rendaku rates in regular compounds. In Irwin (2016a), /y/ and /w/ show slightly higher rendaku rates than expected. In Asai (2014), they show slightly lower rendaku rates than the overall average.

less obstruents. This generalization is backed by the data in (14). The table shows the number of rendaku and non-rendaku surnames with a nasal in E1 in Sugito's (1965) data.

(14) No. of /ta/-surnames by E1-nasal and rendaku (Sugito 1965)

E1 cons.	**t**a	**d**a
/m/	3	33
/n/	0	34

The question to be addressed here is, again, why the type of the last consonant in E1 affects the likelihood of rendaku application in compound surnames, unlike in regular compounds. In both Asai's (2014) and Irwin's (2016a) data of regular compounds, the rendaku rates are not particularly high when E1 is a nasal. Why, then, does rendaku appear to be promoted by E1-nasals in surnames?

As a more general quesion, why do the two kinds of sonorants, approximants and nasals, have different effects on rendaku applicability? The answer will be given in Sections 3.3.5, 3.3.6 and 3.3.7; I will propose that compound surnames are phonologically single stems and that their rendaku patterns are a reflection of stem-internal phonotactics.[24]

One other interesting observation made by Zamma (2005) is that /m/ tends to cause free variation between the rendaku form and the non-rendaku form. Zamma conducted a judgment study in which five native speakers of Japanese were asked to read out surnames written in kanji loud.[25] He notes that speakers tend to vary with respect to rendaku application when E1's last consonant is /m/, especially in surnames with /sawa/ (e.g. u<u>m</u>e-**s**awa ~ u<u>m</u>e-**z**awa; to<u>m</u>i-**s**awa ~ to<u>m</u>i-**z**awa).[26] Taken together with Sugito's (1965) data in (14) where surnames with /ta/ as E2 always undergo rendaku when E1's last consonant is /n/ while allowing a few non-rendaku names when the consonant is /m/, this may suggest that E1-/m/ triggers rendaku less compared to E1-/n/. Particularly

[24]But see the results of a nonce-name experiment in Section 4.3 for further complications.

[25]Although Zamma (2005) does not specifically say what orthography was used, I assume that the surnames were written in kanji, which does not reflect rendaku voicing (see Section 2.2.3).

[26]Zamma (2005:4) writes "It may be that /m/ can sometimes be regarded as voiced in names with *sawa*."

interesting to us here is a question of why there is such a slight difference among the two nasals in their rendaku-triggering effects.

2.3.4 Summary of segmental factors

Japanese compound surnames show complex phonological patterns with respect to rendaku, which are not found in normal compounds. The characteristics of rendaku application in surnames which have been discussed so far based on the findings of previous studies are summarized in (15) below.

(15) A summary of segment-driven rendaku patterns in surnames

- Putative Strong Lyman's Law effects:
  - Voiced obstruents (/z, b, d, g/) in E1 block rendaku
- /k/'s peculiarity:
  - Voiceless obstruents (/s, t, k/) in E1 generally trigger rendaku
  - /k/ exceptionally allows non-rendaku in surnames with /ta/ as E2
- Differences among sonorants:
  - Approximants (/r, y, w/) in E1 tend to inhibit rendaku
  - Nasals (/m, n/) in E1 tend to promote rendaku

2.4 Prosodic factors

This section discusses the prosodic properties of compound surnames in relation to rendaku application in particular. I will show that surnames are different from regular compounds also in terms of prosody and how this interacts with the voicing alternation.

2.4.1 Accent patterns of names

It is documented in the literature that surnames are also somewhat special from the point of view of accentuation. Following Ito and Mester (2016), we may classify words in Japanese into two categories based on the predictability of their general accent patterns. Inflected words such as verbs

and adjectives are either unaccented or have accent on the syllable containing the penultimate or antepenultimate mora (or on the initial syllable for words shorter than three moras). Whether an item is accented or not is lexically determined and therefore unpredictable, but whether the accent is penultimate or antepenultimate is determined strictly based on the type of suffix the item takes. For example, a verb form with the infinitive (or present tense) suffix /-ru/ receives penultimate accent, unless it is lexically specified to be unaccented: e.g. ^{2}[tabé-ru] 'to eat,' ^{2}[tabané-ru] 'to bind'; cf. ^{0}[tame-ru] 'to save.'[27] Descriptively, these suffixes can be considered accent-attracting suffixes, placing accent on the syllable right before. Verb forms with other suffixes such as the past tense /-ta/, on the other hand, receive antepenultimate accent, following the default antepenultimate accent rule (see Martin 1952 for the original proposal; also see Shinohara 2000; Kawahara 2015b; Kubozono 2008), unless they are specified to be unaccented: e.g. ^{3}[tábe-ta] 'eat-Past,' ^{3}[tabáne-ta] 'bind-Past'; cf. ^{0}[tame-ta] 'save-Past.' Thus, for these words, accentedness itself is not predictable, but the location of accent is; if accented, they receive penultimate accent in the case of having a special accent-attracting suffix, otherwise the default antepenultimate accent.[28]

By contrast, for most uninflected lexical items such as nouns, not only accentedness but also accent location is unpredictable. In principle, accent can fall on any syllable, and which syllable gets accented is presumably specified underlyingly.[29] This brings about lexical contrasts in terms of both accent location and accentedness, as can be illustrated by the famous minimal triplet: ^{2}[hási] 'chopsticks,' ^{1}[hasí] 'bridge' and ^{0}[hasi] 'edge.' However, Ito and Mester (2016) imply that proper

[27] As is noted in Section 1.3.3, when pitch accent is under discussion, a superscripted number is placed in front of every word in order to indicate the location/presence of an accented mora, following Ito and Mester (2016). A superscripted '2' indicates that accent is on the penultimate (second-to-last) mora and a superscripted '1' indicates that accent is on the last mora. A superscripted '0' indicates that the word is unaccented (no accented mora).

[28] Kubozono (2008) analyzes verb forms with these accent-attracting suffixes as following the same accent rule as noun-noun compounds, which place accent near the boundary between components. Also see Tanaka, S. and Kubozono (1999); Nishiyama (2010) for more details on the accent patterns of verbs in general.

[29] This is not to say that the grammar of the language should play no role in accent assignment in these words. Statistically speaking, antepenult is the most common accent position (see Kubozono 2008; also note that lexical strata, word length and syllable weight have effects on accentuation). It is conceivable to posit that many words have no underlying accent specification and antepenultimate accent is assigned by the grammar as the default pattern. Other rarer patterns may also be derived by subgrammars (see Ito and Mester 2016) on the assumption that words are somehow tagged with information about which subgrammar applies to them.

nouns form a yet another class of exceptions. They state that three and four mora surnames and place names are either unaccented (e.g. ^{0}[naka-no], ^{0}[hiro-sima]) or have antepenultimate accent (e.g. ^{3}[nága-no], ^{3}[nagá-saki]). Given names can be characterized in the same way; unaccented (e.g. ^{0}[masa-e], ^{0}[taka-o]), else antepenultimate accent (e.g. ^{3}[másа-ko], ^{3}[taká-hiro]).[30] This is unexpected given they are uninflected nouns, which would normally show variability in where accent falls. In other words, surnames and other proper nouns are strange in that the location of accent is predictable (that is, always antepenultimate) unlike in other nouns.[31]

The fact that their accent is always antepenultimate points to another peculiarity of compound surnames; they do not show the accent patterns expected for compounds. Japanese has a complex compound accent rule. Simply put, for compound nouns with a short E2 morpheme (one or two moras), accent generally falls near the boundary between the two elements. Compound accent is thus either on the final syllable of E1 as in (16a) and (16b), or on the initial syllable of E2 (with E2 preserving its original accent) as in (16c). Otherwise, the compound is unaccented.[32]

(16) Compound accent in regular compounds[33]

a.	síro	+	mé	→	^{2}si**ró**-me	'white-eye'
b.	náka	+	yubi	→	^{3}na**ká**-yubi	'center-finger'
c.	migi	+	máe	→	^{2}migi-**má**e	'right-front'

Compound surnames, however, do not necessarily follow this general pattern. As stated above, accent always falls on the syllable containing the antepenultimate mora. Examples are shown in (17).

[30]See Tanaka, S and Kubozono (1999); Sugawara (2012) among others for accent of given names.

[31]Compound surnames are also special in that they do not follow the usual compound accent rule, as we will see below.

[32]For more details, see Kawahara (2015b) and Kubozono (2008) and references therein.

[33]Words may have variable accent patterns. For example, (16c) can also be unaccented according to some speakers.

(17) No compound accent in compound surnames

a.	síro	+	tá	→	^{3}**sí**ro-ta	^{*2}si**ró**-ta	‘white-paddy (surname)’
b.	náka	+	mori	→	^{3}na**ká**-mori		‘center-forest (surname)’
c.	miya	+	máe	→	^{3}mi**yá**-mae	^{*2}miya-**má**e	‘shrine-front (surname)’

In three-mora names with a monomoraic E2 as in (17a), accent is one syllable away from the boundary. Also, even if E2 is a morpheme which usually preserves its original accent (unless the compound is unaccented), as in (17c), the antepenultimate rule still applies. In the case of a four-mora name as in (17b), the accented antepenultimate syllable happens to be the final syllable of E1, and the pattern is what is expected by the compound accent rule. However, given that there is a large number of examples like (17a) and (17c), it is reasonable to conclude that compound surnames do not follow the usual compound accent rule, despite being compounds.

2.4.2 Rendaku-accent correlation

Another issue that has been raised regarding the prosody of compound surnames is an inverse correlation between accentedness and rendaku application. Sugito (1965) points out that compound surnames with rendaku tend to be unaccented while those without rendaku are often accented. Although Sugito’s (1965) data are limited to surnames with /ta/ as E2, Zamma (2001, 2005) and Ohta (2013) also show that surnames with other common E2 morphemes such as /kawa/ ‘river’ exhibit similar tendencies. This is illustrated in (18) by pairs of surnames that are segmentally similar.

(18) Correlation between rendaku application and accentedness in surnames

	No rendaku	Rendaku
a.	^{3}íku-**t**a	^{0}ike-**d**a
b.	^{3}také-**k**awa	^{0}taki-**g**awa
c.	^{3}naká-**si**ma	^{0}naka-**z**ima

In the examples above, the surnames without rendaku are accented (and have antepenultimate accent as expected) whereas those with rendaku are unaccented. Notice that, as in (18c), a

name which is composed of the exact same E1 and E2 morphemes (/naka + sima/ 'center-island') and shows free variation in the application of rendaku may be either accented or unaccented depending on the presence of voicing, as if rendaku and accent are in complementary distribution (^{3}[naká-**s**ima] ~ ^{0}[naka-**z**ima]).[34]

The table in (19) adapted from Sugito (1965) gives the counts of surnames with E2-/ta/ showing rendaku or no rendaku by accentuation. The numbers of surnames which follow the correlation (that is, rendaku names which are unaccented and non-rendaku names which are accented) are shown in boldface. Note that the table includes surnames showing variation in accent (placed under "Variable accent" in the table here) and those showing variation in rendaku (under "Variable rendaku"), unlike the other tables shown above.

(19) No. of /ta/-surnames by rendaku and accentedness (Sugito 1965)

	Accented	Unaccented	Variable accent
No rendaku	**94**	13	10
Rendaku	64	**95**	56
Variable rendaku	8	0	22

The correlation in the data seems to be rather moderate, given that there are a non-negligible number of surnames that go against the generalized pattern (e.g. sixty-four surnames that show both rendaku and accent).[35] Nevertheless, we can still see there is a tendency for /ta/-surnames to be accented if rendaku does not apply and unaccented if it does.[36]

Zamma (2005) also reports the counts of /kawa/-surnames in his data, sorting them by rendaku application and accentuation, as shown in (20) below.

[34] See Tanaka, S. (2005b) for the observation that island names with the morpheme /sima/ show the same kind of rendaku-accent correlation.

[35] See Section 3.4.5 for an explanation about why the correlation is not robust in surnames with /ta/.

[36] It is impossible to calculate an accurate correlation coefficient between rendaku application and accentuation in these data because there is no information available as to how often each surname considered to be variable by Sugito undergoes rendaku/is accented. Note that the rate of variation between rendaku and no rendaku is not the same for all variable surnames.

(20) No. of /kawa/-surnames by rendaku and accentedness (Zamma 2005)

	Accented	Unaccented	Variable accent
No rendaku	**19**	2	0
Rendaku	1	**10**	0
Variable rendaku	0	2	3

Although it may not be a fair comparison due to the smaller data sample, the correlation in /kawa/-surnames seems stronger than that in /ta/-surnames reported in Sugito (1965). Notice that in (20), most names which show rendaku are accented and those which do not show rendaku are mostly unaccented.

Zamma (2005) further looks at surnames with E2 other than /ta/ and /kawa/ and notes that the correlation is highly dependent on the E2 morpheme. Surnames with E2's such as /tani/ 'valley,' /sita/ 'bottom, under' and /ki/ 'tree' follow the pattern to a greater or lesser extent while those with other E2 morphemes such as /saki/ 'cape,' /tuka/ 'mound' and /kuti/ 'entrance, mouth' do not, in that they are almost always accented regardless of rendaku application. Based on this observation, Zamma concludes that the rendaku-accent correlation originally described by Sugito (1965) does not apply to all surnames but still holds true for a non-negligible number of names with particular E2 morphemes, producing a weak tendency in the overall data.

Interesting to note is that this rendaku-accent correlation is characteristic of a particular type of compounds. Besides in compound surnames (see Sugito 1965; Ohta 2013; Zamma 2001, 2005 among others), the same kind of pattern is reported in certain old-fashioned given names containing /sabu-roo/ 'third-son' (Haraguchi 2002), island names with /sima/ 'island'[37] (Tanaka, S. 2005a,b), complex verbs composed of a Sino-Japanese root and /suru/ 'do' (Okumura 1984) and deverbal compounds (Yamaguchi 2011). Notice that these words are either proper nouns or constructions containing inflected verbs, which are described to be alike in terms of the predictability

[37]As we have seen in (18c), the morpheme /sima/ can also appear as E2 in surnames and may exhibit the rendaku-accent correlation. Note, however, that Zamma (2005) classifies /sima/ into items that do not follow the correlation based on the overall patterns of surnames containing the morpheme. It is possible that one same morpheme has different accent and rendaku behaviors in surnames and place names. Zamma (2005:173) also notes that /kawa/ 'river' shows different rendaku behaviors when it appears as E2 in surnames and in river names.

of accent (see Section 2.4.1). By contrast, the rendaku-accent correlation does not seem to exist as a robust pattern in other types of compounds. Alderete (2015) looks at whether there is such a correlation in compounds in general using Rosen's (2001) database of Japanese compound words. Although he finds a trend in the data, it turns out not to be statistically significant.[38] If so, then why do compound surnames exhibit accent patterns that are not seen in regular compounds? Why is the rendaku-accent correlation peculiar to a small class of lexical items including proper nouns?

Before answering these qeustions, we must also address the question of what relatioship holds between rendaku application and accentedness. At this point, it is not obvious which one is the cause of the other, or whether they are in any causal relationship. Does the (non-)application of rendaku affect the accent pattern of the surname? Or does accentedness play a role in determining whether the surname undergoes rendaku? I will show below that answering these questions helps us answer the question raised earlier concerning why the correlation is found in compound surnames but not in normal compounds. In Section 3.4.4, I will propose that the correlation is the result of an interaction between two forces; a compound surname must behave like a single stem accentually, but rendaku application requires compound-like prosodic structures. I will further show that Ito and Mester's (2016) account of unaccentedness in Japanese stems based on metrical feet can be directly extended to compound surnames, thereby addressing the questions of how unaccentedness arises in surnames with rendaku and of why surnames with different E2 morphemes show different degrees of correlation.

2.4.3 The length of elements

As discussed in Section 2.3, Sugito (1965) has pointed out that E1's last consonants affect the rendaku applicability in compound surnames. It should be noted that Sugito's research is only concerned with surnames composed of two-mora E1 and /ta/ 'paddy' as E2.[39] Zamma (2005)

[38] Also note that Rosen's (2001) data include proper nouns such as surnames and place names, which may be contributing to the trend that Alderete notices.

[39] The reason why Sugito (1965) only included such names is most probably that the majority of native Japanese morphemes are bimoraic and /ta/ is by far the most frequent morpheme which appears as E2 in compound surnames

and Asai (2014) conduct larger-scale studies, and find that the length of E1 also affects rendaku in surnames. Generally speaking, names are more likely to undergo the voicing alternation when E1 is monomoraic than when it is bimoraic. This is illustrated in (21) by common surnames composed of one-mora E1 and /ta/. E1's last consonants, which are also word-initial consonants, are underlined.

(21) Monomoraic E1 tends to trigger rendaku

Rendaku:	s̲u-**d**a	t̲o-**d**a	k̲i-**d**a	n̲o-**d**a	w̲a-**d**a	y̲a-**d**a	o-**d**a
No rendaku:	m̲i-**t**a	h̲a-**t**a					

As can be seen, the majority of surnames with monomoraic E1 show rendaku. Recall that approximants (/r, w, y/) in E1's final syllable are characterized as segments that inhibit rendaku in surname (see Section 2.3.3). In (21), however, the blocking effect seems to be overturned by E1's monomoraicity (w̲a-**d**a, y̲a-**d**a).[40]

Zamma (2005) points out that this rendaku-triggering effect of monomoraic E1 is also very much contingent on what morpheme is in the E2 position. Some E2 morphemes such as /kawa/ 'river' behave like /ta/, undergoing rendaku most of the time when preceded by one-mora E1 (e.g. o-**g**awa, sa-**g**awa). Other morphemes such as /tani/ 'valey' and /saka/ 'slope,' however, do not show such patterns (e.g. ko-tani, ta-saka).

We further find a different type of E1-length effect. When E1 is trimoraic, names often undergo rendaku regardless of E1's last consonant. This is illustrated in (22) by common /ta/-surnames with trimoraic E1 morphemes.

(Shirooka and Murayama 2011).

[40]The reader may wonder what happens if E1 is monomoraic and its last consonant is a voiced obstruent, which shows a strong rendaku-blocking effect (see Section 2.3.1). This is hard to test since there are very few native morphemes which begin with a voiced obstruent, due to an independent constraint which bans word-initial voiced obstruents (see e.g. Ito and Mester 2003:211–212). /r/ also does not appear in word-initial position in the native lexicon.

(22) Trimoraic E1 tends to trigger rendaku

Rendaku:	hayas̲i-**d**a	miyak̲o-**d**a	hakam̲a-**d**a	sakur̲a-**d**a	kasiw̲a-**d**a
Variation:	yanag̲i-**t**a ~ yanag̲i-**d**a	yomog̲i-**t**a ~ yomog̲i-**d**a			

As can be seen, when E1 is three moras long, approximants in E1's final syllable, which would otherwise inhibit the voicing alternation, allow the rendaku form (sakur̲a-**d**a, kasiw̲a-**d**a). As for common names, /ta/-surnames with trimoraic E1 always undergo rendaku except for a few which show free variation. These variable names typically contain a voiced obstruent in E1's final syllable (yanag̲i-**t**a ~ yanag̲i-**d**a, yomog̲i-**t**a ~ yomog̲i-**d**a[41]). Recall that voiced obstruents in E1 are argued to block rendaku almost categorically in /ta/-surnames (see Strong Lyman's Law discussed in Section 2.3.1). It is possible that the rendaku-triggering effect of trimoraic E1 and the rendaku-blocking effect of E1-voiced obstruents are both playing a role in these names, causing the variation.

The statement that trimoraic E1 morphemes promote rendaku application may actually be more generalized. As has been mentioned earlier and will be discussed in more detail in the next section, rendaku in surnames is conditioned by E2's lexical propensities in addition to the segmental and prosodic factors we have seen so far; some E2 morphemes are generally more prone to rendaku while others are more resistant. Once we look at such morpheme-specific behaviors of second elements, we find that trimoraic morphemes such as /hayasi/, /taira/ and /tokoro/ frequently undergo rendaku when they appear as E2 in compound surnames. This suggests that rendaku is generally more likely to apply when either E1 or E2 is trimoraic.

Unlike the other characteristics discussed in previous sections, the length effects of compound elements do not seem to be unique to surnames. Irwin (2016a) validates the rendaku-promoting effects of monomoraic E1 in regular compounds. He shows with his large corpus data that the average rendaku rate is higher when the length of E1 is one mora than when it is two moras.

[41] The E1 morpheme /yanagi/ 'willow tree,' which is a type of tree, is claimed to be etymologically a complex word, containing the word /ki/ 'tree' which has undergone voicing (*ya-no-ki 'arrow-shaft-tree' > yanagi; Shinmura 1998). The same is probably true for /yomogi/ 'worm wood.' It is unclear whether ordinary Japanese speakers are aware of the historical morpheme boundaries or how that affects rendaku on the following morpheme.

Tamaoka et al. (2009) also conducted a rendaku judgment experiment to test the effects of E1 length, using nonce words presented as regular compounds. Their stimuli were composed of existing E1 morphemes of various lengths (e.g. /ko/ 'small,' /simo/ 'lower, bottom,' /sakura/ 'cherry tree') and nonce trimoraic E2 morphemes (/hukari/ and /hasuri/). Participants were asked to rate whether presented compounds (e.g. /ko-hukari/, /simo-hukari/, /sakura-hukari/) should undergo rendaku or not. The authors found that the rendaku application rate is higher when E1 is one mora than when it is two moras. (See the paragraphs below for discussion on the comparison between two-mora E1 and three-mora E1.)

The effects of trimoraic E1 in regular compounds has also been discussed in the literature. Rosen (2001, 2003) points out that rendaku applies relatively often when one of the elements is trimoraic. More precisely, Rosen (2001) claims that if either E1 or E2 is three moras or longer, the compound is required to undergo rendaku; if rendaku still does not apply under such a condition, then the E2 belongs to "rendaku-immune" morphemes, which inherently never show the compound voicing.[42] Rosen (2001) calls this condition the "prosodic size factor" and subsequent studies call it "Rosen's rendaku prosodic size rule" (Irwin 2009) or simply "Rosen's Rule" (Irwin 2016b; Vance 2015). Irwin (2009, 2016b) and Vance (2015) show that, although Rosen's Rule is not quite as exceptionless as has originally been claimed, it still captures the general tendencies of rendaku application in large data of normal compounds.[43]

Despite these findings, the psychological reality of Rosen's Rule has been called into doubt. Kawahara and Sano (2014c) conduct a nonce word judgment experiment in order to test whether the length-based rendaku patterns described by Rosen (2001) are internalized in Japanese speakers' minds. Their stimuli are composed of real E1 of either two or three moras (e.g. /mori/ 'forest,' /hayasi/ 'wood') and nonce E2 of three moras (e.g. /semaro/). They predicted that, due to

[42]Rosen (2001:28) originally defines the condition as being when "the second member begins with a voiceless obstruent" (i.e. the compound is eligible for rendaku application), "both members of the compound are of Yamato origin" and "both members of the compound exceed one mora and at least one of the members of the compound exceeds two moras."

[43]Irwin (2009) proposes some modifications to the definition of Rosen's Rule. Interestingly, one of his proposals is to exclude proper nouns, noting "[c]ross-linguistically names often exhibit aberrant behaviour" (Irwin 2014b:188).

Rosen's Rule, the acceptability of rendaku would be higher for nonce compounds with three-mora E1 (/hayasi-semaro/ 'wood-semaro') than for those with two-mora E1 (/mori-semaro/ 'forest-semaro'). Yet the prediction was not borne out; their participants judged rendaku application in both kinds of compounds to be equally acceptable, regardless of E1's length. In Tamaoka et al.'s (2009) experiment with a very similar design (see above), they even found that the rendaku application rate was slightly higher in nonce compounds with two-mora E1 than in those with three-mora E1. These results are at odds with the findings of Irwin (2009, 2016a) and Vance (2015).

What caused such differences between the actual data patterns and speakers' behaviors? It is possible that the patterns seen in the real word data have some diachronic origin, but the rule itself is not synchronically active in Japanese speakers' grammar. Another possibility is that, in the experiments by Kawahara and Sano (2014c) and Tamaoka et al.'s (2009), the rule's effect was masked by confounding factors such as the length of E2. Rosen's Rule, in its original definition, states that rendaku should apply if *one of the elements* exceeds two moras. Since their nonce E2 morpheme is already trimoraic, all the compounds in their stimuli are actually eligible for rendaku by this criterion. This may have caused a ceiling effect, masking potential acceptability differences between compounds with two-mora E1 and those with three-mora E1. Settling the issue would require detailed investigation of the data in these studies and further experimentation, which is out of the scope of this dissertation. Here, I highlight the point that the rendaku-triggering effects of trimoraic elements have been reported in existing regular compounds as well as compound surnames, even though their psychological reality has not been fully proven experimentally.

2.4.4 Summary of prosodic factors

I have shown that rendaku in Japanese compound surnames is affected not only by segmental factors as discussed in Section 2.3 but also by prosodic factors. Some of such factors are found uniquely in surnames and others also play a role in regular compounds. The main points discussed in this section are summarized in (23) below.

(23) A summary of rendaku patterns in surnames with respect to prosody

- An inverse correlation between rendaku and accentuation:
  - ▹ Surnames with rendaku tend to be unaccented
  - ▹ Surnames without rendaku tend to be accented
  - (▹ The correlation is found only in surnames)
- The effects of element length:
  - ▹ Monomoraic E1 promotes rendaku
  - ▹ Trimoraic E1 or E2 promotes rendaku
  - (▹ Both effects are also found in regular compounds)

2.5 Lexical propensities

This section discusses how rendaku in surnames is affected by lexical factors; each E2 morpheme shows an idiosyncratic behavior with respect to rendaku application, and so does each surname.

2.5.1 Lexical propensities of E2

As discussed in Section 2.2.1, rendaku in regular compounds is constrained by E2's lexical propensities to the voicing alternation. Some E2 morphemes are inherently more prone to rendaku than others, and some others are inherently resistant to it (see Nakagawa 1966; Vance 1979, 1980b; Rosen 2001, 2016; Irwin 2016a). Rendaku in compound surnames is similarly affected by the morpheme-specific properties of E2. Kubozono (2005) and Zamma (2005) note that not all E2s behave like /ta/, which shows high sensitivity to the voicing, place, and manner of the consonant in E1's final syllable in terms of rendaku application (Sugito 1965). For example, E2 morphemes such as /kuti/ 'entrance' almost always undergo rendaku[44] while others such as /sita/ 'bottom,

[44]Interestingly, the morpheme /kuti/ is a polyseme and changes its rendaku propensity depending on the meaning. As shown in the examples of regular compounds in (4) in Section 2.2.1, /kuti/ shows much variation in rendaku when it means 'mouth' or metaphorically 'way of speaking.' On the other hand, the morpheme can also metaphorically mean 'taste,' in which case it never undergoes rendaku in compounds: e.g. [ama-kuti] 'sweet-taste'; [kara-kuti] 'hot-taste.' Irwin (2016a:104) calls such morphemes "Jekyll and Hyde elements," and classifies /kuti/ 'way of speaking'

below' very rarely do. Such lexical propensities yield exceptions to the generalizations on rendaku application we have seen above. Surnames with E2-/kuti/ undergo rendaku even when the last consonant of E1 is a voiced obstruent, overriding Strong Lyman's Law, as in [mi̲z̲o-**g**uti] 'trench-entrance.' On the other hand, surnames with E2-/sita/ resist the voicing change even when E1 is trimoraic and E1's final syllable contains a consonant considered to be a rendaku trigger, as in [tutu̲m̲i-**s**ita] 'embankment-bottom.' Also note that even among E2's that are more variable in terms of rendaku application, the propensity to voicing is different for each morpheme. (These morphemes can be called "rendaku lovers," "rendaku haters," and "rendaku waverers" respectively in the terms of Rosen (2001) and Irwin's (2016a).) The segmental factors discussed in Section 2.3 thus interact with such inherent lexical propensities of E2 morphemes, producing complex rendaku application patterns.

E2 morphemes may also have lexically-specific accentuation properties. As pointed out by Zamma (2005), some morphemes such as /saki/ 'cape' cause surnames they occur in to be accented, while others such as /sawa/ 'mountain stream, swamp' make them unaccented. This affects the inverse correlation between rendaku and accent discussed above (Section 2.4.2). Recall that surnames with rendaku tend to be unaccented and those without tend to be accented. The accent specification of E2, however, may override this general pattern. For example, surnames with E2-/saki/ are always accented whether they show voicing or not, as in ^{3}[miyá-**z**aki] 'shrine-cape' and ^{3}[nagá-**s**aki] 'long-cape.' On the other hand, surnames with E2-/sawa/ are almost always unaccented regardless of their rendaku profiles, as in ^{0}[naka-**z**awa] 'center-mountain stream' and ^{0}[naga-**s**awa] 'long-mountain stream'). Thus, surnames that have E2 morphemes with accent specification do not necessarily follow the correlation, unlike those with /ta/ 'paddy' and /kawa/ 'river' as E2.

into rendaku waverers and /kuti/ 'taste' into rendaku immune morphemes. Although not mentioned in Irwin (2016a), it seems that /kuti/ is a rendaku lover or even submissive when it has the third meaning 'entrance' or 'gate' both in surnames and in regular compounds: e.g. [iri-**g**uti] 'enter-gate (entrance),' [de-**g**uti] 'exit-gate (exit).'

2.5.2 Lexical propensities of compounds

What makes the matter even more complicated is that the entire compound surname may also show idiosyncratic behavior regarding the application of rendaku.[45] Recall that /r/ in E1 tends to block the voicing, as I have shown with the examples in (12) above, repeated here as (24).

(24) /r/ in E1 tends to block rendaku (repeated)

a. na̲ri-**t**a ku̲ri-**t**a ku̲ra-**t**a mu̲ra-**t**a mo̲ri-**t**a hi̲ro-**t**a a̲ri-**t**a ...
b. ha̲ra-**d**a ku̲ro-**d**a te̲ra-**d**a

The generalization appears to be true, given that rendaku fails to apply in the majority of common surnames with E1-/r/, as shown in (24a). Notice, however, that there are a few surnames which exceptionally show rendaku, as in (24b). It is virtually impossible to explain why these particular names are exceptions. For instance, there seems to be no clear reason for why rendaku applies in [3][kúro-**d**a], which is segmentally and prosodically very similar to [3][kúri-ta]. If there is any reason, it is simply that the surname /kuro-ta/ 'black-paddy' is lexically specified to undergo rendaku. It will soon be apparent that all of the generalizations about rendaku in surnames that have been discussed so far come with such lexical exceptions. This indicates that the application of rendaku in a surname is determined not only based on phonological factors and lexical propensities associated with its elements but also based on the lexical properties of the name itself. In order to fully account for the rendaku patterns in surnames then, one must take such lexically specific behaviors of compound surnames into consideration.

2.5.3 Summary of lexical propensities

In this section, I have shown that rendaku application in compound surnames is affected by the lexically specific properties of the involved items, in addition to the phonological factors discussed in Sections 2.3 and 2.4, as summarized in (25) below.

[45]This also applies to regular compounds, as discussed in Section 2.2.1.

(25) A summary of the effects of lexical propensities on rendaku in surnames

- Lexical propensities of E2
  - ▹ Each E2 morpheme has a specific propensity for rendaku, which may interact with or override the effects of phonological factors
- Lexical propensities of compounds
  - ▹ Each compound surname has a specific propensity for rendaku, which may interact with or override the effects of phonological factors and E2's lexical propensities

In Chapter 5, I propose a probabilistic grammar equipped with Optimality Theoretic constraints which models the rendaku patterns in surnames. In order to capture the lexically specific patterns discussed above, I propose that some constraints be tagged with each E2 morpheme and also each compound surname, along the lines of Moore-Cantwell and Pater (2016). I will show that a Maximum Entropy grammar (see Goldwater and Johnson 2003) with lexically specific constraints can not only capture the rendaku patterns with lexical exceptions in existing surnames but also model Japanese speakers' productive and variable rendaku application in nonce surnames.

2.6 The phonological status of rendaku in surnames

Thus far, I have laid out issues surrounding the sound patterns of surnames in Japanese, particularly rendaku and accent. One may wonder whether these phenomena should be understood as problems in the synchronic grammar of Japanese, as is claimed throughout this dissertation. In this section, I will consider alternative accounts based on diachrony and sheer lexicalization, and argue that such explanations are untenable. I will first review the history of rendaku application in Japanese, as well as the history of Japanese surnames, and claim that the rendaku patterns seen in current surnames cannot simply be attributed to history. I will also argue against the view that the sound patterns of proper nouns are all lexicalized, in favor of an account in which rendaku application in surnames is a productive process.

2.6.1 History of rendaku application

Rendaku is attested in the oldest substantial texts in Japan, which reflect a variety of the Japanese language spoken by the aristocracy in the capital at the time, Nara, from the early 7th century through the late 8th century (Jodaigo Jiten Henshu Iinkai et al. 1967). The variety is known as Jodai Japanese, or Old Japanese in Western texts (Miyake 2003). It was first suggested by Lyman (1894) and is now widely believed among scholars that rendaku voicing originated from a sound alternation introduced by contracted forms of particles such as genitive *-no* (sometimes transcribed as *-nö* for reconstructed prehistoric Japanese and Old Japanese texts; Miyake 2003) or oblique *-ni*.[46] It is hypothesized that many two-element phrases, which would later be viewed as compounds, had these nasal-initial particles attached to the first element, as in (hypothetical) [yama-no sakura] 'mountain-GEN cherry tree.' After losing the vowel, the particles were reduced to [n], which in turn caused post-nasal voicing (Ito and Mester 1995a,b) of following obstruents, as in [yama-n-sakura] > [yama-n-**z**akura]. Eventually, the nasal segment was further reduced to a voicing feature, and the whole voicing process resulted in a morphophonological phenomenon known as rendaku, which can be seen in Old Japanese, as well as in present-day Japanese, as in the attested form [yama-**z**akura][47] (see the dictionary entries in Jodaigo Jiten Henshu Iinkai et al. 1967:769 for Old Japanese, and Shinmura1998 for present-day Japanese). (For more details on the hypothesis on the development of rendaku from nasal-initial particles, see Lyman 1894; Murayama 1954; Unger 1977; Ito and Mester 2003:99; Vance 1982, 2005a.)

Although the phenomenon of rendaku has existed for a long time, there have been several changes in terms of its application. One such example is a change in the domain of the categorical rendaku blocking factor, namely, Lyman's Law. Diachronically, the ban on multiple voiced obstru-

[46] Alternative explanations have also been proposed. See e.g. Kindaichi 1938 (via Kindaichi 1976/2005); Murayama 1954; Yamaguchi 1988.

[47] A more accurate transcription of the Old Japanese form may be [yama-^{n}(d)zakura], given that word-medial voiced obstruents were prenasalized in Old Japanese (see Hamada (1952); Martin (1987); Frellesvig (2010); also see Vance et al. (2014); Miyashita et al. (2016) for dialects of present-day Japanese which have retained prenasalized obstruents). Prenasalization of voiced obstruents may also have contributed to the loss of the preceeding nasal segment: [yama-n-^{n}(d)zakura] > [yama-^{n}(d)zakura].

ents applied in the domain of a prosodic word instead of a stem. Unger (1977) claims that, in Old Japanese (early 7c.–late 8c.; see above), rendaku was blocked if either E1 or E2 contained /z/, /b/, /d/ or /g/, since voicing E1's initial consonant would create multiple voiced obstruents within the *whole word* (e.g. /mîd̲u/ + /t̲öri/ → mîd̲u-**t**öri, *mîd̲u-**d**öri 'water-bird'; Vance 2005c:31; the transcription is his).[48] The commentary section of *Jidaibetsu Kokugo Daijiten Jodaihen* (Jodaigo Jiten Henshu Iinkai et al. 1967), the definitive dictionary of Old Japanese, also has a similar statement: "[i]t seems that rendaku was inhibited when the last consonant of the first element [of a compound] was a voiced obstruent" (ibid:31). Vance (2005c) and Vance and Irwin (2013) confirm the validity of this generalization by scrutinizing headwords in the aforementioned dictionary. They find that, although not many relevant and reliable datapoints are available,[49] Old Japanese compounds containing a voiced obstruent in E1 indeed resisted rendaku application with only a few exceptions.[50] Unger (1977) refers to this ban on multiple occurrences of voiced obstruents within the whole compound as the "strong version of Lyman's Law," and later studies simply call it "Strong Lyman's Law," the term we have already seen.

Other studies suggest that the domain of the law has changed over time, and its effect has become "weaker" in later stages of the language. Sakurai (1972) observes that, in Early Middle Japanese, spoken between the late 8th century and the late 12th century (or during a time known as

[48]Unger (1977) makes his argument referring to the descriptions by Ramsey and Unger (1972) and Miyake (1932), and the latter in turn attributes the original discovery to *Kogen seidaku kō* (Ishizuka 1801) by the Japanese philologist Ishizuka Tatsumaro (1764–1823).

[49]Vance (2005c) finds sixty-two compounds with a voiced obstruent in E1, and Vance and Irwin (2013) deal with even smaller thirty-nine compounds because of stricter exclusion criteria. The scarcity of their data is due to the inherent underrepresentation of (underlying) voiced obstruents in Old Japanese and ambiguities or inconsistencies in the transciprtion of voicing in some words.

[50]Vance (2005c) and Vance and Irwin (2013) report four such exceptions, and they all have a voiced obstruent in a position earlier than the final syllable of E1. The only example given by Vance and Irwin (2013) is [mad̲ara-**b**usuma] 'multicolor-bedding' composed of /mad̲ara/ and /p̲usuma/. The other three Strong Lyman's Law violators described by Vance (2005c), which are excluded from Vance and Irwin (2013), are all compounds composed of three morphemes, where the blocker (a voiced obstruent) is in the first element and the rendaku site (an initial voiceless obstruent) is in the third element: e.g. /kuz̲u/ + /pa/ + /k̲ata/ → kuz̲u-pa-**g**ata 'kudzu-leaf-vine'). Although the data are too scarce to make any conclusion, this may suggest that the effects of Strong Lyman's Law were weaker when the blocker and the potential rendaku site were farther apart (Vance and Irwin 2013) and/or when there were multiple morpheme boundaries between them.

the Heian Period starting with the relocation of the capital from Nara to Kyoto), Strong Lyman's Law was still playing a role as a rendaku blocking factor but already allowed a number of exceptions. This contrasts with normal (stem-bounded) Lyman's Law, which had categorical effects in the language at the time according to Sakurai (1972), just as it does in present-day Japanese. Endo (1980) argues that the force of Strong Lyman's Law saw a sharp decline around the late-9th century, based on his observation of documents from the Heian Period (794–1185) and the Kamakura Period (1185–1333).[51] Toda (1988) also examines whether the presence of a voiced obstruent in E1 affected rendaku application in later periods. She conducts corpus studies using two old dictionaries of Japanese: *Nippo Jisho* (Jesuit missionaries 1603/1604), a Japanese-Portuguese dictionary compiled by Portuguese missionaries, which reflects the pronunciations of Late Middle Japanese spoken from the late 12th century through the 16th century, and *Waei Gorin Shūsei* (Hepburn 1867/1872),[52] the oldest Japanese-English dictionary which reflects the pronunciations of Early Modern Japanese spoken from the 17th century through the 19th century. Toda concludes that voiced obstruents in E1 did not exert any influence on the rate of rendaku in either period. As mentioned above (see Section 2.3.1), it has been proposed that the word-bounded law still has a weak effect in present-day Japanese (see e.g. Kindaichi et al. 1988:264; Sato 1989). Yet recent studies show that such a tendency is not found in large corpora of rendaku in regular compounds (Irwin 2014a, 2016a; Sano 2015; cf. Asai 2014 and the discussion in Sections 2.3.1 and 2.3.3) and that there is no experimental evidence for its psychological reality (Kawahara and Sano 2014c).

The table in (26) summarizes the effects of Strong (word-bounded) Lyman's Law for regular compounds in the Japanese language in each period of time, based on the findings of the literature. Note that normal (stem-bounded) Lyman's Law has been active throughout the development of the language.

[51] Endo (1980) defines the law as a ban on a sequence of two voiced obstruents across a morpheme boundary, rather than a general ban on multiple voiced obstruents within a word. He thus does not look at cases where a voiced obstruent occurs in a syllable earlier than the final syllable of E1.

[52] Toda does not specify which edition is used.

(26) History of Strong Lyman's Law effects for regular compounds[53]

Time Period	Strong Lyman's Law
Old Japanese (7c.–late 8c.)	Active
Early Middle Japanese (late 8c.–late 12c.)	Active? Many exceptions
Late Middle Japanese (late 12c.–16c.)	Not active
Early Modern Japanese (17c.–19c.)	Not active
Modern/present-day Japanese (20c.–present)	Not active

Ito and Mester (2003) interpret this language change as constraint rerankings in Optimality-Theoretic terms (Prince and Smolensky 1993/2004). They posit two kinds of OCP(voice) constraints: one which bans the occurrence of multiple voiced obstruents in the domain of a stem,[54] and the other which does so in the domain of a word. The two constraints were both ranked high in Old Japanese, but the word-bounded constraint was demoted and ranked below rendaku triggering constraints sometime in the period of Early Middle Japanese. As a result, rendaku is no longer blocked by voiced obstruents in E1 in the later stages of the language, including present-day Japanese.[55]

The loss of Strong Lyman's Law is not the only change that the language has undergone. The literature shows that there are other differences between Old Japanese and present-day Japanese in terms of rendaku application. Vance and Irwin (2013) point out that, in Old Japanese, other than

[53]The table ignores possible dialectal differences. For example, Old Japanese and Early Middle Japanese refer to varieties spoken in the Kinki region (mainly Nara and Kyoto), which was the political center of the country at the time, but Modern/present-day Japanese is based on the variety spoken in the Kanto region (Tokyo). The patterns in Late Middle Japanese are also inferred from Toda's (1988) observations of words in the oldest Japanese-Portuguese dictionary, which may have been influenced by a variety spoken in Nagasaki, where trades with foreign contries were conducted.

[54]To be precise, Ito and Mester (2003) use the term "morpheme" as the domain of the constraint.

[55]Another possible interpretation is that Japanese only had stem-bounded OCP(voice) as an active constraint throughout its history and Old Japanese had a different way of analyzing compounds in morphophonology. Assuming that compounds composed of multiple stems can be recursively a stem (e.g. $[A]_{stem}$ + $[B]_{stem}$ → $[[A]\text{-}[B]]_{stem}$), as will be proposed for compound names in this dissertation (see Section 3.2.2), stem-bounded OCP(voice) can take effect over compounds. In this view, what changed is not constraint rankings but the way compounds are interpreted in grammar. Here, I leave this question open, since testing the predictions of the two analyses empirically or giving a formal analysis of the language change between Old Japanese and Early Middle Japanese is beyond the scope of this dissertation. The reader is also referred to Section 3.3.1 for related discussion. I thank Junko Ito for bringing my attention to the issue and discussion.

voiced obstruents, the liquid /r/ in E1 also inhibited rendaku. In their data compiled from the Old Japanese dictionary mentioned above, compounds with /r/ in E1's last syllable have a remarkably lower average rendaku application rate (26%) when compared to the overall average (41%). This is different from the pattern in regular compounds in present-day Japanese; according to Irwin's (2016a) corpus study, the rate of rendaku application with /r/ in E1's final syllable (69.8%–74.3%) is only slightly lower than the overall average (74.3%–76.9%)[56] (also see Section 2.3.3).

The attentive reader may have noticed that the rendaku patterns in Old Japanese, which are no longer seen in regular compounds in present-day Japanese, are similar to the patterns found in current surnames; they are both subject to the rendaku-blocking effects of voiced obstruents and /r/ in the first element. Does this simply mean that surnames are old, and thus, their peculiar rendaku patterns are a reflection of the sound patterns in Old Japanese? Before answering the question, I will review the history of Japanese surnames, discussing how they arose and how long they have existed.

2.6.2 History of Japanese surnames

According to historical studies on Japanese surnames (see Toyoda 1971; Kato 1984; Okutomi 2004; Sakata 2006 among others), Japan has a long history of having names for the groups people belong to, besides names for individuals. In ancient times, the society was made up of kinship groups or clans called *uji* 氏. In Classical Japan (592–1185), the country was ruled by the Imperial Court, and the Emperor granted his vassal clans official names known as *ujina* 氏名 (lit. 'clan name') or *sei* 姓.[57] Sometime in the Heian period (794–1185), many aristocratic clans started using an additional name referring to their immediate family and close relatives in order to differentiate

[56]Irwin (2016a) shows two kinds of rendaku rates; the first one is based on the entire data of his corpus and the second one is based on compounds with high word frequency.

[57]This clan name was also referred to simply as *uji* or *shi* (the Sino-Japanese reading of the kanji 氏), or as *honsei* 本姓. It seems that there were no substantial meaning differences between these terms (Hora 1966:157; Sakata 2006:23,25). Besides such clan names, clans were also awarded *kabane* 姓. Even though it was written with the same kanji as *sei* 姓, *kabane* was more of a noble title based on their profession and rank, and was different from a clan name (Hora 1966:157; Sakata 2006:22–23; Okutomi 2004:15).

themselves from other families of the same clan (who had the same clan name). These additional names, which I refer to here as "family names," were often taken after the place of their residence.[58] For example, a nobleman of the clan *Taira* 平 living in a place called *Kajiwara* 梶原 had the clan name *Taira* and the family name *Kajiwara*. Initially, such family names were used for practical purposes and were not passed down to posterity. It was very common for a father and his son or brothers from the same family to have different family names if they were living in different places. Towards the end of the Heian period (794–1185) and in the Kamakura period (1185–1333), however, the system of patrilineal inheritance was developed and each family's sense of identity became stronger. As a result, offspring (usually the first son) began succeeding to the family name of his father, along with his father's properties.

The warrior class also developed their own family names in a similar way. In the Heian period, many warrior clans moved to the country, for example *Togoku* ('the Eastern countries' now known as the Kanto region), to reclaim land. When they were settled in the region for generations, they started using family names on top of their clan names. Family names were often taken after the name of the place which their ancestors had first cultivated and created their base in. In the Muromachi period (1336–1573), a family name was already an official and important property of each family line, and many kept it even if they moved to a new region afterwards. Once family names gained such an official status, they became known as *myoji* 名字/苗字.

Later in the Meiji era (1868–1912), every citizen in Japan was required to have one name for their family unit (see below). For practical purposes, I call these names under the new system "surnames," as opposed to "clan names" and "family names." Many people who were originally of the aristocratic or warrior class adopted their family names (that is, *myoji*) as their surnames, and fewer used their original clan names (*ujina* or *sei*). Thus, many surnames we see today can be traced back to family names from Medieval Japan (late 12c.–16c.), or clan names from Classical Japan (6c.–late 12c.). This fact is also reflected in the language today. In present-day Japanese, the words *shi* 氏 (the Sino-Japanese reading of the kanji for *uji*), *sei* 姓 and *myoji* 名字/苗字 are all

[58]Some also used the name of their profession, the temple which their ancestor had built or the place of their secondary house in the suburbs of Kyoto.

used to refer to "surnames" with some differences in formality. (For more details on the history of clan names, family names and current surnames, see Hora 1952, 1966; Toyoda 1971; Kato 1984; Okutomi 2004; Sakata 2006.)

We have seen how aristocrats and warriors, who constituted the upper class and the upper-middle class in Japanese society at the time, came to have names for their clans and families, which would later become their surnames. What about the commonalty, who made up the majority of the society? Did they also have family names (*myoji*), and if so, since when? It is well known that, in the Edo period (1603–1868), most ordinary people were not allowed to have official family names. Like clan names that were given by the Emperor, family names were now a privilege of established social classes, associated with the possession of a territory. This fact has led people in later eras, including scholars of Japanese culture and folklore, to believe that historically, commoners had no family names (see e.g. Yanagita 1933/1998, 1946/1998; Kindaichi 1965). The view is reconsidered by recent historians, however. Hora (1952, 1966) claims that peasants and merchants did have family names but they simply could not (or did not) use them officially. Hora's studies show informal records from the Edo period listing ordinary people's family names and given names. Subsequent studies report that similar documents are found in several parts of Japan, and some even suggest that commoners were using their family names in private ceremonies and religious rituals (Tamura 1953; Hayashi 1953a,b; also see Toyoda 1971:157–168). More recent studies further show that common people already had family names before the Edo period. Sakata (2006:44–48) argues that the use of family names was spread to areas near Kyoto before or around the 15th century, based on his observations of documents from the Muromachi Period (1336–1573). Many of the ordinary people's family names reported in Sakata (2006) are ones we see quite frequently as surnames today, such as *Arai* 新井 'new-well,' *Hayashi* 林 'woods,' *Tanaka* 田中 'paddy-center' and *Eguchi* 江口 'bay-entrance.'[59] These names were probably taken after place names, created based on the landscape or characteristics of their place of residence or modeled after already existing family names of aristocrats or warriors.

[59]Note that the pronunciations of the names might be actually different from what we see today.

In the Meiji era (1868–1912), the new government declared equal rights for all its citizens, and dramatic reforms of the social system were implemented. The reforms included the enactment of the Surname Law (the Meiji Government 1875), which required all citizens to have a surname.[60] The law made people of the commoner background register one name for their family unit as their official surname.[61] It also stated that those who do not have a family name (*myoji*) handed down from their ancestors must create a new surname on their own. There are a number of real or anecdotal stories about how people had to invent surnames because of the law (see e.g. Hora 1952:412(2); Yanagita 1946/1998:577). However, historians now agree that such cases were rather rare (see Hora 1952, 1966; Toyoda 1971:171–172 among others). As discussed above, common people already had family names around the 15th century (Sakata 2006), and seem to have kept them even in the Edo period (1603–1868), even though they were not allowed to use them officially. In the new era, they simply registered their family names, which had been passed down from generation to generation, as their surnames under the new system.

To summarize, even though the current system of surnames is relatively new, Japanese people have long had the tradition of having names for their kinship groups besides their given names. Aristocrats and warriors had the names of their clans and families from the Classical times (6c.–late 12c.) and the Medieval times (late 12c.–16c.) respectively. Commoners are also believed to have had names for their families beginning in the 15th century.

2.6.3 Not merely a historical vestige

We have seen that both rendaku and surnames have a relatively long history. We have also seen that the rendaku patterns in Old Japanese compounds are quite similar to those in current surnames. Again, the question at stake here is why surnames show patterns that look peculiar when compared

[60]Before this law, another law on surnames, which "permitted" all citizens to have a surname, had been enacted in 1870 (the Meiji Government 1870). Not many people registered their surnames then, however, possibly because of the long custom of not using family names officially (Okutomi 2004:148–150).

[61]The new system also required aristocrats and warriors to renounce many of their privileged rights. As stated above, they could only register one "surname," and thus had to give up their clan names and family names in the traditional sense.

to regular compounds in present-day Japanese. A simple answer could be that surnames are older than regular compounds, and the former have retained the sound patterns of Old Japanese.

This diachronic explanation, however, is not satisfactory for several reasons. First, although some surnames based on clan names can indeed be as old as Old Japanese (7c.–late 8c.), not all surnames we see today existed in the late 8th century, the time when the language was spoken. As mentioned above, in the Meiji era, many people registered their family names handed down from Medieval Japan (late 12c.–16c.). It is possible that people in the Medieval times made their family names after existing words such as place names, which had Old Japanese traits including Strong Lyman's Law. However, it is documented that many surnames were newly invented around the 12th and 13th centuries. Examples come from the family names of one of the oldest aristocratic clans *Fujiwara* /huzi-hara/ 藤原 'wisteria-field.'[62] Many subgroups of the *Fujiwara* clan moved from Kyoto to other regions of the country in the 12th and 13th centuries. Some of them served as executive officers of the region or became warriors and reclaimed land. Once they were settled, many derived their family names by combining part of their clan name and the name of their residence or profession. For example, a *Fujiwara* clan in the country of *Ise* invented the name *Ito* /i-too/ 伊藤 by combining /i/, the first syllable of *Ise*, and /too/, the Sino-Japanese reading of the morpheme *fuji* /huzi/ from *Fujiwara*. Similarly, *Sato* /sa-too/ 佐藤 is composed of /sa/, the first syllable of the name of their profession *Saemon no Jo* 'a kind of government official,' and /too/.[63] Others also combined part of their clan name and geographical words, as in *Fujita* /huzi-ta/ 藤田 'wisteria-paddy,' *Fujikawa* /huzi-kawa/ 藤川 'wisteria-river,' and *Fujisaki* /huzi-saki/ 藤崎 'wisteria-cape.' Notice that these names exhibit Strong Lyman's Law effects, with a voiced obstruent in E1 blocking rendaku ([huzi-ta], [huzi-kawa], [huzi-saki]). Recall also that Strong Lyman's Law was already inactive in the 12th century, as shown in (26) (also see Toda 1988). If these trun-

[62]See Asai (2014); Vance and Asai (2016) for a description of the alternation /h/ → [w] seen in the pronunciation of [huzi-wara] derived from /huzi-hara/.

[63]There are many other names derived in the same way from *Fujiwara*. Examples include (but are not limited to) *Saito* 斎藤, *Kato* 加藤, *Kudo* 工藤, *Sudo* 須藤, *Endo* 遠藤, *Naito* 内藤 and *Ando* 安藤. Interestingly, the morpheme /too/ undergoes rendaku quite often to become [**d**oo], even though it is a Sino-Japanese morpheme which is expected to be resistant to the voicing alternation.

cated compound surnames were created around then, why do they still show systematic rendaku blocking effects of E1-voiced obstruents? Simply stating that surnames are old and thus have kept the sound patterns of Old Japanese does not explain the behaviors of surnames which were coined after the time of Old Japanese. The facts suggest that there was something special about surnames which created the peculiar but systematic patterns that are still seen today.

Secondly, proper nouns are subject to historical change like common nouns, and it is unlikely that surnames can simply retain old sound patterns as is. The reading of a surname, including the voicing of a consonant, could be changed relatively easily until recently. This is mainly due to the fact that the Japanese household register known as *koseki*, which is a type of civil registry for each family unit, only uses the kanji script to represent registered people's surnames and does not provide their readings.[64] We find records (official and unofficial) stating that some people changed their surnames from a rendaku form to a non-rendaku form or vice versa. For example, the founder of the music instrument company Yamaha Corporation, *Torakusu Yamaha* (1851–1916) changed his surname (underlyingly /yama-ha/ 山葉 'mountain-leaf') from *Yamaba* with rendaku voicing to *Yamaha* with no rendaku (Miura 2012). Another example of a change in rendaku in names also comes from the founder of a famous Japanese corporation. The founder of Toyota Motor Corporation is *Kiichiro Toyoda* (1894–1952) (the surname is underlyingly /toyo-ta/ 豊田 'rich-paddy'). The company was originally named *Toyoda* after his actual surname, but it was later changed to *Toyota* without rendaku for purposes of euphony and because of superstitions.[65] These anecdotes suggest that Japanese speakers are fairly flexible with the variation in rendaku in names. On the Internet, we can also find stories on how people's close ancestors or they themselves have changed the readings of their surnames. Since the reading of a surname could be changed at

[64]Nowadays, most Japanese citizens also possess other types of official identification documents such as passports and resident cards. These documents provide the readings of their surnames in scripts other than kanji, whih can indicate rendaku voicing. For this reason, it is no longer easy to have the official reading of one's surname changed.

[65]The website of Toyota states "[i]t has been regarded as a favorable transition from 'Toyoda' to 'Toyota,' because voiceless consonants sound more appealing than voiced consonants. In addition, through the concept of 'jikaku' (counting the number of strokes in writing characters to determine good and bad luck), its eight-stroke count is associated with wealth and good fortune. [...]" (Toyota Motor Corporation 1995-2017b; also see Toyota Motor Corporation 1995-2017a).

each person's will, surnames were also subject to historical change. Proposing that surnames have retained Strong Lyman's Law from Old Japanese, one must still explain why change in the law's domain was systematically blocked specifically in surnames with a voiced obstruent in E1, even though a change in rendaku application could have happened and it did happen in other contexts. If it is true that old surnames with a voiced obstruent in E1 have kept their non-rendaku forms since the time of Old Japanese, there must be some factor, besides them being just old, which caused them to be particularly resistant to the change.

Relatedly, studies show that not all sound patterns seen in surnames are a reflection of old sound patterns. As mentioned above, /r/ in E1 inhibits rendaku in surnames as well as in Old Japanese compounds. The other approximants /y/ and /w/ in E1, however, do not behave alike in the two word classes; they are rendaku-blockers in surnames (see (13) in Section 2.3.3), but they seem to promote rendaku in compounds in Old Japanese (Vance and Irwin 2013). Another difference is the behavior of E1-/m/; it promotes rendaku in surnames (see (14) in Section 2.3.3), but rather tends to block it in Old Japanese (Vance and Irwin 2013). Again, these small differences suggest that the sound patterns in current surnames are not merely carried over from Old Japanese.

In summary, historical facts and phonological facts both indicate that the peculiarity in rendaku application in surnames cannot be viewed as a historical vestige.

2.6.4 Productivity and predictability

Another important reason why rendaku in surnames should be dealt with by a sychronic grammar is that its application is a productive process. One may think that speakers memorize the sound patterns of proper nouns, including rendaku voicing. This may be true when it is concerned with the names of individuals that people know in person. However, there are situations in which Japanese speakers have to determine whether a surname undergoes rendaku or not. As I stated earlier (Section 2.2.3), surnames are usually written in the kanji script, which does not reflect rendaku voicing. It happens quite often that a Japanese speaker encounters the written form of the name of someone he or she does not know. The rendaku judgment is easy if it is a very common surname which

shows consistent rendaku application behavior, but it may also be a name that shows variation in rendaku, as in [naka-ta] ~ [naka-da] or [yama-saki] ~ [yama-zaki], or a rare name that most people have never seen but is a potential rendaku undergoer, as in /sori-ta/ 剌田 'splash-paddy' or /tubu-saki/ 粒崎 'grain-cape.' In such a case, a speaker must judge the applicability of rendaku in each surname, without any prior memorization. In Section 4.3 below, I will further show the results of nonce name experiments, suggesting that Japanese speakers' judgments are not random but based on phonological factors. I claim that, if speakers apply rendaku to unknown or newly coined surnames using such "phonological knowledge," the phenomenon must be understood as a productive process explained by phonological grammar.

Another point that should be made about the sound patterns of surnames is its regularity. As we have seen, rendaku in surnames shows much lexical variation, but phonological (segmental and prosodic) factors also play an important role in its application. This seems different from rendaku in regular compounds, where phonological factors are often overridden by lexical factors (see e.g. Irwin 2016a; Rosen 2016). In other words, rendaku in surnames is more predictable than rendaku in regular compounds. Their accent patterns are also regular. Recall that accentuation in surnames (and proper nouns in general) is limited to two possibilities: unaccented or penultimate accent (see Section 2.4.1). This is in stark contrast to other uninflected nouns which show much variation in the location of accent. Again, this suggests that surnames are less affected by lexical specification compared to other nouns. Following the basic and common assumption in Generative Phonology that predictable patterns are regulated by grammar, I claim that the accent patterns of surnames are governed by the phonological grammar of Japanese.

To summarize, the sound patterns of surnames show both productivity and predictability, indicating that they must be accounted for by phonological theory.

2.6.5 Summary

In this section, I have reviewed the history of rendaku and the history of Japanese surnames, and have argued that the peculiar rendaku patterns in current surnames are not simply the reflection of

the sound patterns in Old Japanese. I have also pointed out the role of phonological grammar in rendaku and accent in surnames. Many aspects of the phenomenon are productive and predictable, and hence they should be explained in terms of the theory of phonology.

2.7 Chapter summary

In this chapter, I have described the rendaku and accent patterns of compound surnames and discussed their differences from those of regular compound words. One of the main characteristics of rendaku in surnames is that its application is conditioned by the features of the last consonant of the first element (E1). For example, we find the effects of so-called Strong Lyman's Law, whereby a voiced obstruent in E1 inhibits rendaku voicing (see Sugito 1965 among others). Surnames also show peculiar prosodic patterns. They do not follow the usual compound accent rule and the location of accent is predictable (see Ito and Mester 2016). Furthermore, whether a surname has accent or not is inversely correlated with the presence of rendaku voicing (see Sugito 1965; Zamma 2005; Ohta 2013 among others). I have presented a brief overview of the historical development of rendaku and Japanese surnames, and argued that the peculiar sound patterns of surnames should be accounted for by phonological grammar.

CHAPTER 3

Proposal and analyses: Rendaku and accent under stem phonology

3.1 Chapter overview

This chapter presents an account of the phonological patterns of Japanese surnames layed out in the previous chapter. In Section 3.2, I make a proposal that compound surnames are represented as stems in the grammar of Japanese and are subject to the phonology of stems. Section 3.3 shows that the peculiar rendaku patterns in surnames can be derived from the application of stem-internal phonological restrictions. Section 3.4 shows that their prosodic characteristics also follow from the application of stem phonology. Section 3.5 gives short remarks on the other issues such as lexical propensities and how they should be addressed in phonological theory.

3.2 Proposal

3.2.1 Non-compositionality of surnames

As we have seen in Chapter 2, most Japanese surnames have a compound structure, but they behave differently from common noun compounds in terms of accentuation and rendaku application. Apart from phonology, surnames also have special characteristics in the domain of semantics; despite being compounds, their meanings are not compositional. For example, the surname /ta-naka/ is composed of /ta/ 'paddy, rice field' and /naka/ 'center, middle.' Etymologically speaking, it presumably derived from a place where inhabitants were surrounded by paddies. The name itself,

however, does not actually mean 'the center of paddies'; rather, it refers to an individual. By contrast, regular compound words such as /mati-naka/ 'city-center' and /yo-naka/ 'night-center (time around/after midnight)' are semantically compositional, having a meaning derived from combining the meaning of each element. As will be discussed below, findings in the literature suggest that regular compound words may also have non-compositional meanings and the dichotomy between compositional compounds and non-compositional compounds is not always absolute. Nonetheless, a crucial difference between surnames and regular compounds is that the former are never compositional and show the greatest degree of non-compositionality.

It has been argued that phonotactics, morphological parsing and semantics interact with each other. Hay (2003) shows that probabilistic phonotactics plays an important role in morphological parsing, serving as a cue to morpheme boundaries.[1] For example, English speakers find the morpheme boundary more easily in the prefixed word *inhumane* than in *insincere*, because the /nh/ transition in the former is unlikely to be found within a simplex word, and thus it facilitates the processor to posit a boundary. She proposes a dual-route model of morphological processing, where a morphologically complex word is accessed either as a whole (or via a "direct" or "whole word" route in her terms) or through its component parts (or a "decomposed" route). She then argues that a word like *dislocate* is likely to be accessed via a whole word route since it contains relatively weak phonotactic boundary cues and is less decomposable. (Note that the /sl/ transition across the morpheme boundary is well attested morpheme-internally, as in *slow*.) She further claims that the way a word is accessed affects the representation of that word, bringing about further consequences in other domains. Less decomposable words such as *dislocate* are more likely to be accessed and stored as a whole word or one unit as opposed to decomposed parts in speakers' mental lexicons. This can in turn cause semantic drift, making the word acquire semantically opaque and non-compositional meanings. Hay (2003) provides evidence for the hypothesis with nonce word experiments and a corpus study investigating the relationship between English words' phonotactic properties and the transparency of their dictionary definitions.

[1] Also see Hay et al. (2004). For the role of phonotactics in speech segmentation in general, see e.g. Saffran et al. (1996a,b), and other references cited in Hay (2003).

Hay's (2003) study has implications for proper nouns which are morphologically complex, such as compound surnames in Japanese. Hay demonstrates that phonology can affect semantics through morphology; complex words with phonological characteristics of simplex words tend to be represented as one whole word, which further derives non-compositional meanings. Extending the claim that the three linguistic domains interact with one another, I propose that semantics can also affect phonology through morphology. More particularly, the intrinsic non-compositional meanings of proper nouns affect their representations and pronunciations; if a complex word has inherently no semantic compositionality as in the case of a proper noun, it will be represtend as one whole word, which can in turn derive simplex-word-like phonological patterns. I will show below that such a process in fact takes place in the parsing and phonological realization of Japanese compound surnames. I will claim that compound surnames are represented as single stems in morphophonology and that their peculiar rendaku and accent patterns stem from such representational properties.

Before turning to the details of the proposal, I will present one study suggesting that compound surnames are treated as single units by Japanese speakers. The famous study by Miller (1956) points out that the number of objects an average person can hold in their working memory is about seven. Miller also argues that humans can group pieces of information into one meaningful unit or a chunk, and that each chunk counts as one object in memory. Suzuki (2016) conducts informal experiments with Japanese speakers, testing Miller's hypothesis.[2] He presents numbers (e.g. 1, 2, 3), alphabets (e.g. a, b, c) and Japanese compound surnames (e.g. *Nishimura* /nisi-mura/ 'West-village') to college students and tests how well they recall the stimuli. The results replicate Miller (1956), showing that the participants remember about seven objects on average. He also finds that they perform similarly with all the three kinds of stimuli. That is, one number, one letter and one surname are all treated as a chunk in the recollection task. This indicates that Japanese speakers treat a compound surname as one unit, even though they are composed of two morphemes and also

[2]I thank Shigeto Kawahara for bringing my attention to the study.

of several phonological elements.[3]

In the next subsection, I make a proposal on the morphophonological representations of Japanese compound surnames and compare them to those of regular compounds.

3.2.2 Recursive stems and stem-phonology

I have shown that Japanese compound surnames are complex words with inherent non-compositionality. I argue that Japanese speakers access them directly as whole words rather than through their components in morphological processing, and store them as single units in their lexicons. In order to formalize the idea, I propose that compound surnames, which are composed of stems, are also labeled as stems in morphology. In other words, compound surnames are recursive stems, which have a compound structure but also behave like a single stem (functioning as a single personal name in semantics). This contrasts with regular compounds; a regular compound is also composed of two stems, but the output does not form a stem itself. The representations of compound surnames and regular compounds are illustrated in (27) below.

(27) Representations of compound surnames and regular compounds

a. A compound surname: A stem composed of stems
$[\,A\,]_{stem} \quad + \quad [\,B\,]_{stem} \quad \rightarrow \quad [\,[\,A\,]_{stem}\text{-}[\,B\,]_{stem}\,]_{stem}$

b. A regular compound: Two combined stems
$[\,A\,]_{stem} \quad + \quad [\,B\,]_{stem} \quad \rightarrow \quad [\,A\,]_{stem}\text{-}[\,B\,]_{stem}$

I further claim that these representational differences bring about phonological consequences; compound surnames are subject to the phonology of stems while regular compounds are exempt from such stem-phonology and further goes through phonological operations for compounds. I will show below that the peculiar rendaku and accent patterns of compound surnames can be attributed to the application of stem-phonology.

[3]Of course, this study by itself does not provide evidence that compound surnames are different from regular compounds. It is possible that speakers also treat a regular compound word as one unit.

3.3 Analysis: Deriving segmental factors

I have argued that compound surnames are treated as single stems and that they are subject to phonological restrictions on stems rather than compounds. In this section, I will show that the application of stem-phonology derives the segmental factors in rendaku application in surnames discussed earlier, such as so-called Strong Lyman's Law and the peculiarity of /k/.

3.3.1 Stem-bounded Lyman's Law blocks rendaku: The law is not strong

The proposal that a compound surname is represented as a stem in the grammar of Japanese explains why rendaku in surnames is blocked by a voiced obstruent in the first element (E1) unlike in regular compounds. It is not that the effect of Lyman's Law extends beyond the stem-level, as is implied by the oft-used term "Strong Lyman's Law." Compound surnames are stems (with an internal compound structure), and so the normal stem-bounded law applies. Rendaku, which would create multiple voiced obstruents *within a stem*, is thus blocked, as exemplified in (28).

(28) Stem-bounded Lyman's Law applying to the whole compound surname

[si͟ba]$_{stem}$	+	[t͟a]$_{stem}$	→	[[si͟ba]-[**t**a]]$_{stem}$	*[[si͟ba]-[**d**a]]$_{stem}$
[huz͟i]$_{stem}$	+	[t͟a]$_{stem}$	→	[[huz͟i]-[**t**a]]$_{stem}$	*[[huz͟i]-[**d**a]]$_{stem}$

The proposed analysis assumes only one active OCP(voice) constraint in Japanese phonology. Stem-bounded OCP(voice) applies to regular compounds, yielding the rendaku blocking effect known as Lyman's Law (e.g. /hana + taba/ → hana-ta͟ba, *hana-**d**a͟ba 'flower-bunch'). The same constraint applies to compound surnames, banning the occurrence of multiple voiced obstruents not only within the second element (e.g. /taka + su͟gi/ → taka-su͟gi, *taka-**z**u͟gi 'tall-cedar') but also within the whole name (e.g. /huz͟i + ta/ → huz͟i-ta, *huz͟i-**d**a 'wisteria-paddy'). In other words, we need not posit word-bounded OCP(voice) or different constraint rankings for regular words and surnames.[4] The same grammar can derive the effects of both normal Lyman's Law and Strong

[4]See Section 2.6.1 for Ito and Mester's (2003) account of the difference between Old Japanese, which shows Strong Lyman's Law effects, and present-day Japanese, in which only normal Lyman's Law is operative, based on constraint

Lyman's Law. The difference between the two types of compounds arises from the difference in their representations in morphophonology.

The grammar is illustrated in the tableaux in (29), which exemplify cases where rendaku applies or fails to apply in surnames and regular words. For the sake of simplicity, I ignore lexical or within-word variation here, and posit that the constraint RENDAKU, which requires voicing of the initial consonant of the second element of a compound, outranks IDENT(voice), which penalizes a voicing mismatch between the input and the output in correspondence. RENDAKU should be understood as a bundle of constraints that trigger the voicing alternation, such as REALIZE-MORPHEME and IDENTITYAVOIDANCE, which will be discussed in more detail below. It is outranked by stem-bounded OCP(voice), which prohibits the occurrence of multiple voiced obstruents in the domain of a stem (even if that stem itself constrains stems).

(29) i. Rendaku application and non-application in surnames

[[hana]$_{stem}$ + [ta]$_{stem}$]$_{stem}$	OCP(voice)	RENDAKU	IDENT(voice)
(a) hana-ta		*!	
☞ (b) hana-da			*

[[huz̲i]$_{stem}$ + [ta]$_{stem}$]$_{stem}$	OCP(voice)	RENDAKU	IDENT(voice)
☞ (c) huzi-ta		*	
(d) huzi-da	*!		*

ii. Rendaku application and non-application in regular compounds

[huz̲i]$_{stem}$ + [tana]$_{stem}$	OCP(voice)	RENDAKU	IDENT(voice)
(e) huzi-tana		*!	
☞ (f) huzi-dana			*

[hana]$_{stem}$ + [ta̲ba]$_{stem}$	OCP(voice)	RENDAKU	IDENT(voice)
☞ (g) hana-taba		*	
(h) hana-daba	*!		*

rerankings.

As can be seen, candidate (b) [hana-**d**a] with rendaku application is optimal for the compound surname /hana-ta/ 'flower-paddy (surname)' due to the ranking RENDAKU ≫ IDENT(voice). On the other hand, for /huz̲i-ta/ 'wisteria-paddy (surname)' with a voiced obstruent in E1, rendaku is blocked as can be seen in the winning candidate (c) [huz̲i-ta]. Voicing the initial consonant of E2, as in (d) *[huz̲i-**d**a], would create two voiced obstruents within the whole name, which itself is a stem, and the high-ranked stem-bounded OCP(voice) eliminates the candidate. This is in contrast to the regular compound /huz̲i-tana/ 'wisteria-trellis' which undergoes rendaku, as in (f) [huz̲i-**d**ana], despite having a voiced obstruent in E1. Notice that the whole compound is not a stem in its morphological representation, and thus stem-bounded OCP(voice) is not relevant. In the case of regular compounds, the constraint penalizes a candidate with rendaku voicing only if it already has a voiced obstruent in E2, as in (h) *[hana-**d**ab̲a] for the input /hana-tab̲a/ 'flower-bunch.'

In short, under the proposed analysis, Strong Lyman's Law effects, which are often considered to be peculiar characteristics of surnames in present-day Japanese, are nothing but the application of normal Lyman's Law, which is also found elsewhere in Japanese phonology. Strong Lyman's Law itself does not need any special theoretical treatment, and the difference between surnames and regular words in terms of the rendaku blocking condition can be derived from one and the same grammar.

3.3.2 Identity Avoidance triggers rendaku: /k/ is not peculiar

Stem-internal phonotactics can not only block rendaku but also trigger rendaku. Japanese has gradient root/stem consonant cooccurrence restrictions based on place and voicing identity, which will be referred to here as "Identity Avoidance."[5] Using data compiled from *Kōjien* (Shinmura 1998), a large dictionary of comtemporary Japanese, Kawahara et al. (2006) show that stems containing consonants of the same place in adjacent syllables are generally underrepresented in the native Japanese lexicon. For instance, words with two labial sounds such as /maba/ (unattested) are sta-

[5]Avoidance of similar consonants is found in a number of languages. See e.g. Yip (1998), Walter (2007) and references therein.

tistically rarer than words with labial and coronal sounds such as /mada/ 'yet.' Kawahara et al. (2006) also show that voicing contributes to the cooccurrence restrictions in the case of obstruents. Stems with adjacent pairs of homorganic voiceless obstruents such as /s...s/, /t...t/ and /s...t/ are rarer compared to those with homorganic obstruents that disagree in voicing such as /s...z/, /t...d/ and /s...d/. I have proposed that compound surnames are treated like stems, and thus are subject to stem-internal phonology. If Identity Avoidance is a property of stems, surnames should also disfavor the occurrence of similar consonants in adjacent syllables.

I argue that dispreferences for sequences of homorganic voiceless obstruents within stems are indeed reflected in the way rendaku applies in surnames. Recall that surnames with /ta/ 'paddy' as E2 always undergo rendaku if E1's last consonant is /s/ or /t/, as shown in Sugito's (1965) data in (11), repeated here as (30).

(30) No. of /ta/-surnames by E1-voiceless obstruent and rendaku (Sugito 1965)

E1 cons.	**t**a	**d**a
/s/	0	35
/t/	0	26
/k/	13	31

Deriving a surname by combining an E1-morpheme with /s/ or /t/ in its last syllable (e.g. /a̲sa/ 'hemp,' /ma̲tu/ 'pine tree') and E2-/t̲a/ with initial /t/ creates voiceless coronal obstruents in adjacent syllables underlyingly (e.g. /a̲sa-t̲a/, /ma̲tu-t̲a/). Notice that, assuming that compound surnames are stems, such consonant sequences constitute disfavored *stem-internal* homorganic obstruent sequences. Application of rendaku, which changes the voicing of one of the obstruents, can repair the marked structure. As mentioned above, /s...z/, /t...d/ and /s...d/ with voicing contrasts are preferred over /s...s/, /t...t/ and /s...t/. Rendaku then serves as a way to dissimilate an identical or similar consonant sequence within a stem in terms of voicing. In other words, rendaku is triggered by stem-internal phonotactics, specifically, Identity Avoidance. This is illustrated in (31).

(31) Rendaku triggered by stem-internal Identity Avoidance

[a͟sa]$_{stem}$	+	[t͟a]$_{stem}$	→	[[a͟sa]-[**d**a]]$_{stem}$	*[[a͟sa]-[**t**a]]$_{stem}$
[ma͟tu]$_{stem}$	+	[t͟a]$_{stem}$	→	[[ma͟tu]-[**d**a]]$_{stem}$	*[[ma͟tu]-[**t**a]]$_{stem}$

Under this analysis, the fact that rendaku occurs less often in E2-/ta/ when it is preceded by E1-/k/ can be explained (see Section 2.3.2 above). Since voicing E2's initial consonant does not repair a marked structure, rendaku applies less often (e.g. /ak͟i/ + /t͟a/ → [aki-ta], *[aki-**d**a] 'autumn-paddy').[6] The peculiarity of /k/ is then simply the result of the promotion of rendaku in surnames having an underlying /s...t/ or /t...t/ sequence by Identity Avoidance and the lack of such rendaku-triggering effects in surnames having an underlying /k...t/ sequence, which is already dissimilar in place.

The proposed analysis also makes new predictions about how rendaku should apply in surnames with E2's other than /ta/. If Identity Avoidance is at work, /k...k/ sequences should also trigger rendaku, just like /s...t/ and /t...t/ sequences. That is, surnames having a /k/-initial morpheme such as /kawa/ 'river' as E2 are expected to undergo rendaku when E1's last consonant is also /k/. Impressionistically, we indeed find many examples of surnames with a /k...k/ sequence that show rendaku (with some possible variation), as in (32a), and fewer examples that do not, as in (32b).

(32) Rendaku in /kawa/-surnames with /k/ in E1's final syllable

a. na͟ka-**g**awa to͟ku-**g**awa hu͟ka-**g**awa ka͟ke-**g**awa ta͟ki-**g**awa (~ta͟ki-**k**awa)
b. ta͟ke-**k**awa se͟ki-**k**awa yo͟ko-**k**awa

In Section 4.2, I will show that the prediction is actually borne out in a larger scale corpus study.

To summarize the facts about /k/ in E1's final syllable, under the hypothesis that surnames are subject to stem-internal phonotactic restrictions such as Identity Avoidance, it is no longer

[6]Note, however, rendaku may still apply in surnames with a /k...t/ sequence. I argue that voicing is also promoted by REALIZE-MORPHEME, which requires a compound to undergo rendaku. See Section 3.3.8 for the details. See Chapter 5 for how to capture lexical variation.

strange that /ta/-surnames with /k/ in E1 do not undergo rendaku as often as those with /s/ or /t/. Since E1-last /k/ and E2-initial /t/ do not form an identical or similar sequence of consonants, there is not much need for them to dissimilate in voicing. Surnames with E1-last /k/ and E2-initial /k/, on the other hand, are expected to show rendaku because of the effects of Identity Avoidance.

3.3.3 Identity Avoidance as a blocker: Labial cooccurrence restriction

The hypothesis that rendaku in surnames is governed by stem-internal place cooccurrence restrictions makes one other prediction; avoidance of multiple labial consonants within a stem will function as a rendaku blocker. As is mentioned earlier (see Section 2.2.1), when E2-initial /h/ undergoes rendaku, it becomes [b], changing not only in voicing but also in manner and place (e.g. /osi + hana/ → [osi-**b**ana] 'pressing-flower').[7] This alternation creates a case of place cooccurrence if a surname has E1 with /m/ in the final syllable (e.g. /ume/ 'plum') and E2 with initial /h/ (e.g. /hara/ 'field'). That is, in the case of surnames with /h/-initial E2, rendaku can cause a problem, instead of solving one, from the perspective of the Identity Avoidance requirement. Although there are not so many common surnames that fit the description, two are shown in (33).

(33) Rendaku creating a labial-labial sequence [m...b] in surnames

a.	ume	+	hara	→	ume-**h**ara ~ ?ume-**b**ara	'plum-field'
b.	ima	+	hasi	→	ima-**h**asi ~ ?ima-**b**asi	'present-bridge'

As indicated by "~," these names variably undergo rendaku, and according to the corpus data presented in Section 4.2, the non-rendaku forms are more common than the rendaku forms.[8] Ac-

[7]The /b/-/h/ alternation can be traced back to the historical sound change */p/ > /ɸ/ > /h/ (Ueda 1898 via Takayama 2012; also see Kiyose 1985; Miyake 2003:66–77,164–166). Even though /h/ itself is no longer a labial sound in present-day Japanese, it still behaves as if it is the voiceless counterpart of /b/ in certain phonological phenomena, including rendaku.

[8]Other common surnames with the /m-h/ sequence include /sima-hukuro/, /ma-huti/, and /yama-he/, and they all usually undergo rendaku. Although this may suggest the lack of labial cooccurrence restriction effects in surnames, the above names all have additional rendaku triggering factors, such as trimoraic E2, monomoraic E1 and so-called rendaku lover E2, respectively. The general scarcity of data also makes it hard to tease apart different possible explanations. See the results of corpus studies and more discussion in Section 4.2.4.2.

cording to Sugito (1965), nasals in E1's last syllable generally trigger rendaku (see Section 2.3.3).[9] The fact that the names in (33) tend to resist rendaku can be attributed to avoidance of a 'would-be' labial-labial sequence. The hypothesis that place cooccurrence restrictions play a role in rendaku application in surnames thus predicts that the alternation occurs less often with E2-initial /h/ when E1's last consonant is /m/, in comparison to, for example, when E1's last consonant is /n/.[10]

3.3.4 Notes on Identity Avoidance in regular compounds

Before turning to other segmental factors in rendaku, some notes on regular compounds and Identity Avoidance are in order. It has been proposed in the literature that Identity Avoidance is also operative in rendaku in regular compounds. However, research to date has not reached a consensus as to what kind of role it plays, how much of a role it plays, or whether the effect is even significant.

Kawahara and Sano (2014b) argue that Identity Avoidance is a rendaku triggering factor in regular compounds. They test the claim by conducting a nonce word judgment experiment, where Japanese speakers are asked to judge rendaku applicability in nonsense compounds such as /ika-kaniro/, which underlyingly have identical moras across the morpheme boundary, and those such as /ika-taniro/, which do not have such identical moras. The results show a weak but significant effect of avoidance of moraic identity; that is, by and large, Japanese speakers like to apply rendaku more in the former type than in the latter ([ika-**g**aniro] > [ika-**d**aniro]).

However, Kawahara and Sano (2014b) only test cases where compounds have a sequence of "identical moras" (e.g. /ika-kaniro/), and do not include cases where compounds have just an "identical consonantal sequence" (e.g. /ika-keniro/). It is known that Japanese speakers have

[9]Note, however, that the results of a corpus study as well as a nonce-name epxeriment reveals that, when overall data are concerned, E1-nasals do not promote rendaku as much as in surnames with /ta/ as E2. See Section 3.3.5 for dicsussion. See Sections 4.2.4.2 and 4.3.4.2 for the corpus and experimental results.

[10]The reader may be wondering whether there could be a sequence of other labial sounds through rendaku application, as in [b...b] and [p...b]. Notice that if a surname has /b/ in E1's last syllable, as in /siba-hara/, rendaku is blocked anyway by (Strong) Lyman's Law, and it is hard to examine the effect of [b...b] avoidance independently. It is also difficult to test the effect of [p...b] avoidance; except in loans, the distribution of /p/ is limited to certain phonological contexts in present-day Japanese (see the sound change of */p/ discussed above), and it almost never appears as the last consonant of E1.

stronger dispreferences for moraic identity over consonantal identity (see Kawahara and Sano 2014a, 2016, who test the existence of Identity Avoidance as a rendaku blocking factor in regular compounds, as mentioned below). It is then possible that avoidance of identical consonantal sequences, which I claim to be at work in rendaku in compound surnames (see Section 3.3.2 above), has only a weak or non-significant effect on rendaku application in regular compounds. We should also note that Kawahara and Sano's (2014b) experiment uses nonce words, and that the effect of Identity Avoidance as a trigger is not yet fully confirmed with existing words.[11] In fact, an analysis of existing compound words in the rendaku database compiled by Irwin and Miyashita (2013-2016) suggests that Identity Avoidance only has a very limited role in rendaku in regular compounds. (See Section 4.2.5 for details.)

Other studies claim that Identity Avoidance also functions as a rendaku blocker (Sato 1989; also see Takayama 1992, Labrune 2012). Sato (1989:255) states "[r]endaku tends to be avoided when it would create two identical or similar sounds in a row." (The translation is mine.) His examples include the following compounds: /tobi + hi/ → [tobi-hi], *[tobi-**bi**] 'jumping-fire'; /siage + kanna/ → [siage-kanna], *[siage-**g**anna] 'finishing-plane (tool).' According to Sato's (1989) claim, rendaku is blocked in these words since the alternation would create a sequence of two identical syllables with voiced obstruents, as in [bi-bi], or two identical voiced obstruents in adjacent syllables, as in [ge-ga]. These cases of rendaku non-application can be characterized as a combination of Identity Avoidance and Strong Lyman's Law, given that it is E1's last syllable containing a voiced obstruent which inhibits the creation of another syllable containing the same voiced obstruent (and possibly the same vowel) in the compound.

There is a debate as to whether such a synergetic effect is real. Irwin (2014a, 2016a) refutes Sato's (1989) claim, showing that Identity Avoidance and Strong Lyman's Law combined do not exert significant effects on the applicability of rendaku in a large set of regular compounds (Irwin

[11] Asai (2014) reports the rendaku rates of compounds with identical consonantal sequences in his magazine-based corpus (see Chapter 2), but he does not specifically test the effect of Identity Avoidance by comparing them to the rates of compounds with non-identical consonantal sequences. Toda (1988:86-87) states that the rendaku rates of compounds with identical moras are slightly higher than the overall average in Late Middle Japanese (12c.-16c.) and Early Modern Japanese (17c.-19c.) based on her observations of dictionary entries (see Section 2.6.1 for the details of the study), but does not test the effect of consonantal identity.

and Miyashita 2013-2016; Irwin et al. 2017). On the other hand, Kawahara and Sano (2014a, 2016) conduct a series of experimental studies and show that Japanese speakers do exhibit dispreferences for rendaku application when both factors are present, as described by Sato (1989). In their experiments, participants are presented with nonsense compounds such as /iga-keniro/ and /aza-keniro/, and are asked whether voicing should occur or not. The results show that speakers find rendaku application undesirable in the former type as compared to in the latter ([iga-**g**eniro] < [aza-**g**eniro]). Note that both compounds have a voiced obstruent in the first element, but [iga-**g**eniro] with a sequence of identical voiced obstruents is particularly disfavored.[12] Kawahara and Sano conclude that Identity Avoidance, when combined with Strong Lyman's Law, plays a role as a rendaku-blocking factor in regular compounds.[13] The discrepancy between Irwin (2014a, 2016a) and Kawahara and Sano (2014a, 2016) raises an interesting question: How do Japanese speakers' behaviors in linguistic tasks exhibit phonological patterns that are not present in the lexicon? Settling the issue is beyond the scope of this dissertation (see Kawahara and Sano 2014a, 2016 for discussion and their interpretation). I simply note here that a combination of Identity Avoidance and Strong Lyman's Law is claimed to inhibit rendaku in regular compounds, but whether the effects are valid, or how such patterns arise in Japanese speakers' minds if they do at all, still remains arguable.

Lastly, it has also been proposed that a ban on multiple labial sounds in a stem plays a role as a rendaku-blocking factor in normal compounds (Akinaga 1977a; also see Nakagawa 1966; Kindaichi 1976/2005; Kawahara et al. 2006; Suzuki 2015; Kumagai 2016, 2017). It is observed that /h/-initial stems containing /m/, such as /hama/ 'shore,' /hime/ 'princess' and /himo/ 'cord,' tend to resist rendaku application when they appear as the second element of a compound. Akinaga (1977a) argues that rendaku is blocked in such compounds since the alternation would create a

[12]Kawahara and Sano (2016) also find that Japanese speakers apply rendaku even less if it would create identical moras in a row, as in [iga-**g**aniro] derived from /iga-kaniro/. They conclude that avoidance of moraic identity has stronger effects than avoidance of consonantal identity.

[13]Kawahara and Sano (2014c) specifically test the effect of Strong Lyman's Law but find no clear evidence. In their participants' judgments, [iga-**z**emaro] (derived from /iga-semaro/) with Strong Lyman's Law violation is no worse than [aka-**z**emaro] (derived from /aka-semaro/) without a voiced obstruent in E1. Combining the results from the other papers, they conclude that Strong Lyman's Law is not strong enough to show any effects (in regular compounds) unless it is combined with Identity Avoidance.

sequence of two labial sounds, namely, [b...m], as shown in (34).[14]

(34) Rendaku inhibited due to labial cooccurrence in regular compounds

a.	suna	+	hama	→	suna-hama	*suna-**b**ama	'sand-shore'
b.	mai	+	hime	→	mai-hime	*mai-**b**ime	'dancing-princess'
c.	kutu	+	himo	→	kutu-himo	*kutu-**b**imo	'shoe-lace'

Kawahara et al. (2006) maintain that the non-application of rendaku in these words stems from the place cooccurrence restrictions discussed above (see Section 3.3.2); stems with labial consonants in adjacent syllables are disfavored in the lexicon. The statement can be translated into a constraint like *[lab]...[lab]$_{\text{stem}}$, which bans the occurrence of multiple labial sounds within a stem. It should be noted that the restriction is not a categorical rendaku blocking factor, and there are compounds which undergo rendaku despite creating a labial-labial sequence (e.g. /koi-humi/ → [koi-**b**umi] 'love-letter'; /beta-home/ → [beta-**b**ome] 'complete-admiration'[15]). Nonetheless, the effect of the constraint is also attested in certain experimental settings (see Kumagai 2016, 2017), suggesting that the patterns are psychologically real.

It is important to note that the Identity Avoidance effects in rendaku in regular compounds discussed in this subsection are not all the same in terms of their domain. In the case of labial cooccurrence restriction exemplified in (34), compounds would have two labial sounds *within a single stem* (i.e. E2) through rendaku application. On the other hand, experimental stimuli used by Kawahara and Sano (2014a,b, 2016) testing the effect of Identity Avoidance either have an underlying sequence of identitical voiced obstruents, or would have such a sequence by rendaku application, not within one stem but *across the morpheme-boundary* between E1 and E2 (e.g. /ika-kaniro/; /iga-keniro/). If the place cooccurrence restrictions described in Kawahara et al. (2006) are only active within stems, the prediction is that (would-be) identical sequences across a

[14]Akinaga (1977a) claims that there also used to be an effect of cooccurrence restrictions on nasality in this case. It is argued that Japanese voiced obstruents were historically prenasalized Hamada (1952); Martin (1987); Frellesvig (2010). Rendaku which voices and prenasalizes the /h/ would create a sequence of two nasal(ized) sounds (e.g. [suna-$^{\text{m}}$bama]).

[15]Note that the example [beta-**b**ome] also has a voiced obstruent in E1. /beta/ is one of the rare native roots starting with a voiced obstruent.

morpheme boundary will not affect rendaku application in regular compounds. That is, Identity Avoidance will not promote nor inhibit rendaku if the two consonants to be dissimilated appear in E1 and E2, as in /ika-kaniro/ and /iga-keniro/, whereas it will have an effect if they both appear in E2, as in the case of labial cooccurrence restriction such as /suna-hama/. By contrast, in the case of a compound surname which is represented as a single stem, such stem-internal Identity Avoidance is expected to affect rendaku application whether the consonants are separated by the E1-E2 boundary or not, as discussed in Sections 3.3.2 and 3.3.3.[16]

To summarize, previous studies discuss the role of Identity Avoidance in rendaku in regular compounds, but its existence or the domain of such an effect has not yet fully been confirmed. The analysis I have proposed predicts that boundary-spanning Identity Avoidance will have effect on rendaku application only in compound surnames and not in regular compounds. In Section 4.2, I will present the results of corpus studies which support the claim.

3.3.5 E1-nasals: Lack of Identity Avoidance and expanded stem-phonology

I have shown that, in compound surnames with an obstruent in E1's last syllable, stem-internal phonological restrictions such as OCP(voice) and Identity Avoidance affect rendaku application. I now turn to surnames with a nasal consonant in E1's last syllable. I will show that the rendaku patterns in these names can also be explained in terms of stem-phonology; a lack of Identity Avoidance effects causes moderate rendaku application, but some expanded versions of stem-internal constraints may promote rendaku in some (but not all) surnames.

Sugito (1965) points out that surnames with /ta/ as E2 and a nasal consonant in E1's final syllable mostly undergo rendaku, as shown in (14) in Section 2.3.3, repeated here as (35).

[16]Notice that this is parallel to the case of Lyman's Law or stem-bounded OCP(voice). Lyman's Law is active within a stem, and only a voiced obstruent in E2 blocks voicing of E2-initial obstruent in regular compounds. For surnames, on the other hand, the law applies to the whole compound, and a voiced obstruent in E1 blocks rendaku.

(35) No. of /ta/-surnames by E1-nasal and rendaku (Sugito 1965)

E1 cons.	**t**a	**d**a
/m/	3	33
/n/	0	34

However, an extensive corpus study of surnames reveals that E1-nasals do not particularly promote voicing (at least not as much as it has been suggested by Sugito (1965)) in surnames with E2 morphemes other than /ta/. When the data of existing surnames are taken as a whole, the average rendaku application rates of names with a nasal in E1 are only slightly higher than the overall average (see Section 4.2.4.2 for details and complications). Furthermore, the results of a nonce-name experiment testing the productivity of rendaku application in surnames do not find the effects of E1-nasals (see Section 4.3.4.2). Taking these corpus and experimental results into consideration, I argue that surnames with an E1-nasal undergo rendaku only moderately due to an inapplicability of Identity Avoidance. That is, /m/ or /n/ in E1's last syllable never forms a similar or identical consonant sequence with a following E2-initial obstruent, and voicing is neither promoted or inhibited in terms of Identity Avoidance, producing variation between rendaku and non-rendaku within surnames with an E1-nasal (e.g. /si<u>m</u>a-kawa/ → [si<u>m</u>a-kawa] 'island-river'; cf. /i<u>m</u>a-kawa/ → [i<u>m</u>a-**g**awa] 'present-river').

The question still remains as to why rendaku seems to be triggered almost categorically by E1-nasals particularly when the second element is /ta/ in the data of existing surnames. I argue that it is an extended version of a stem-internal phonological operation, namely, post nasal voicing, which promotes rendaku application in those names. It is well known that the constraint POST-NASALVOICING or *NT, which requires obstruents to be voiced after a nasal in the same stem, is an active constraint in native Japanese phonology (Ito and Mester 1995a,b; cf. Rice 1997). In native roots, voiceless obstruents never appear after a moraic nasal (e.g. ✓[tombo] 'dragonfly' but not *[tompo]).[17] In the past tense formation of a verb, the past tense morpheme /ta/ surfaces as [da] after a nasal (e.g. /si<u>n</u>-ta/ → [si<u>n</u>-**d**a] 'die-Past,' /no<u>m</u>-ta/ → [no<u>n</u>-**d**a] 'drink-Past'; cf. /si<u>r</u>-ta/

[17] See Kawahara (2002) for an account of exceptional cases based on Output-Output Faithfulness.

→ [sit-ta] ‘know-Past’, /ne-ta/ → [ne-ta] ‘sleep-Past’). These static patterns and alternation patterns can both be characterized as instances of the application of *NT, the ban on a stem-internal sequence of a moraic nasal and a voiceless obstruent.[18] The patterns are, however, still different from the case of rendaku application in surnames discussed above, where an onset nasal of E1’s last syllable triggers voicing of E2’s initial obstruent across a vowel (e.g. /yama-ta/ → [yama-**d**a]).[19]

I propose that there is a constraint which prohibits a voiceless obstruent after a syllable with an onset nasal, namely, *NVT, an extended version of *NT, and it targets at surnames with /ta/. It may seem to be an ad hoc constraint with no clear phonetic motivation[20] and with not much effect in the actual lexicon as can be seen in many roots with the banned structure (e.g. /anata/ ‘you’; /mata/ ‘fork/crotch’). However, we do find one case where *NVT derives a voicing alternation. The adjectival suffix /-(a)sii/ may surface as [-(a)zii] when it is attached to a word that has a nasal in its last syllable (e.g. [mutum-a**z**ii] ‘intimate,’ [susam-a**z**ii] ‘awful,’ [himo-**z**ii] ‘starving’, [imi-**z**i(i)] ‘terrific’; cf. [subar-asii] ‘wonderful,’ [osoro-sii] ‘horrible’) (see Kindaichi 1976/2005; Matsuura 1996).[21] To the best of my knowledge, no other suffixes than /-(a)sii/ (and related /-(a)siku/) show this kind of voicing alternation, and it even has quite a few exceptions (e.g. [yakam-asii] ‘noisy’; [sami-sii] ‘lonely’). I propose that /ta/ is particularly sensitive to the constraint *NVT, and voices

[18] I assume that the derivational suffix /ta/ is part of the verb stem.

[19] Surnames may also have E1 which ends in a moraic nasal. In such a case, rendaku applies almost categorically (e.g. /sen-ta/ → [sen-**d**a] ‘thousand-paddy’), unless the second element is “rendaku immune” or a “rendaku hater.”

[20] See Hayes and Stivers (1996) for the phonetic grounding of the constraint *NT.

[21] More precisely, the suffix can be analyzed as underlyingly /-(a)sik/, and attachment of another adjectival suffix /-i/ (with double-marking of adjetivization) causes the deletion of /k/, resulting in [(a)si-i]. If /-u/ is attached, it functions as an adverb, as in [susam-azik-u] ‘awfully’ and [osoro-sik-u] ‘horribly.’ By definition, /-(a)sik/ is a suffix which forms an adjective by attaching to a (possibly reduplicated) noun root or a verb root, with the meaning of ‘-like’ or ‘-y,’ and it has existed since the time of Old Japanese (see Jodaigo Jiten Henshu Iinkai et al. 1967; Nihon Kokugo Daijiten Dainihan Henshū Iinkai 2000). In present-day Japanese, there is a large number of adjectives ending in [(a)sii], which were originally formed by suffixation of /-(a)sik-i/. Although the process of word-formation is not very productive any more, new adjectives may still be coined with the suffix, often with the intention of making the derived word sound somewhat funny (e.g. [kaigai-sii] ‘foreign country-y’; [aho-aho-sii] ‘stupid-stupid-y’). Treating the initial /(a)/ of the suffix as an epenthetic vowel, we may give a derivational account of post-nasal voicing by *NT involving opacity, without resorting to the effect of *NVT: e.g. /susam/ + /sik-i/ → susam-**z**ik-i (post-nasal voicing by *NT) → susam-**az**ik-i ([a]-insertion). However, the analysis does not work for cases where rendaku applies even though the preceding vowel is part of the root and is not [a], as in [imi-**z**i(i)] ‘terrific.’

when preceded by a syllable with an onset nasal. Though stipulative, it is necessary to account for why E1-nasals trigger rendaku specifically in surnames with /ta/ as E2.[22]

Another issue about E1-nasals raised by the data in Sugito (1965) and Zamma (2005) is that the rendaku triggering effect of /m/ is less strong than that of /n/ (see Section 2.3.3). The generalization turns out to be true in a corpus of existing surnames; the average rendaku rate is generally higher for surnames with E1-/n/ than for those with E1-/m/ (Section 4.2.4.2). (Note, however, that the effect is not found in a nonce-name experiment; see Section 4.3.5 for discussiion on why there is such a disprepancy between the corpus and the experiment). The pattern holds when E2 begins with a voiceless obstruent other than /h/, i.e., /s/, /t/, or /k/, suggesting that it is not simply the application of the labial cooccurrence restriction on E1-[m] and E2-[b] derived from /h/ discussed above (see Section 3.3.3). Why do the two kinds of nasals, /m/ on one hand and /n/ on the other, affect voicing of E2-initial obstruents in a slightly different manner? I propose an account based on stem-internal phonology with some historical quirk.

Let us first review a case of historical sound change which has been claimed to affect rendaku application in regular compounds. It is documented that some instances of /b/ have undergone a sound change to become /m/ (Unger 2004),[23] and that certain E2 morphemes containing /m/ derived from /b/ never undergo rendaku (see Nakagawa 1966:313–314; Vance 1987:147; Irwin 2014a, 2016a; Vance and Asai 2016). This is presumably because rendaku was always blocked by Lyman's Law in their original form with /b/. For example, /kemuri/ 'smoke' was originally /keburi/, and voicing never occurred in a compound with this morpheme as E2, since it would have

[22]The reason for why /ta/ in particular shows such patterns remains unclear. Interestingly, the effect of *NVT is also found in rendaku in some place names. The Sino-Japanese word /kyoo/ 'capital' is usually resistant to voicing, but may undergo rendaku when appearing in the names of districts in Kyoto. It voices when E1's last consonant is a nasal, as in [kami-**g**yoo-ku] 'upper-capital-ward' and [simo-**g**yoo-ku] 'lower-capital-ward,' or when the consonant is /k/, as in [naka-**g**yoo-ku] 'center-capital-ward,' by Identity Avoidance; otherwise, rendaku does not apply, as in [sa-**k**yoo-ku] 'left-capital-ward' and [nisi-**k**yoo-ku] 'West-capital-ward.'

[23]The nasal /m/ and the stop /b/, which was realized as prenasalized [^{m}b] when voiced obstruents had prenasalization (Hamada 1952; Martin 1987; Frellesvig 2010), were interchangeable in many words in Early Middle Japanese (Martin 1987:31–32; Unger 2004:331–332). The same kind of variation can still be found in certain words in present-day Japanese, as in [sabisii]~[samisii] 'lonely' and [samui]~[sabui] 'cold,' often with one of the variants being considered colloquial or regional.

created two voiced obstruents within a stem (e.g. /suna-keburi/ → [suna-keburi], *[suna-**g**eburi] 'sand-smoke'). It seems that some of these E2 morphemes retained their status as being "rendaku immune" even after the sound change, and they continue to resist rendaku still in the language today (e.g. /suna-kemuri/ → [suna-kemuri], *[suna-**g**emuri]).

A similar story can account for the behavior of surnames with /m/ in E1. In the case of a surname, if /b/ appears in E1, rendaku fails to apply due to the application of Lyman's Law over the whole compound. /b/ in some E1 morphemes turned into /m/, but those morphemes retained the "rendaku blocker" status as their idiosyncratic trait, and lowered the average rendaku triggering rate of surnames with E1-/m/. Given this diachronic explanation, one could propose that a constraint like *mVD, which bans the occurrence of [m] and a voiced obstruent within a stem, has some gradient effect. I will discuss the issue of whether this constraint is really present in the synchronic grammar of Japanese when I present the results of a productivity experiment in Section 4.3.[24]

In summary, generally speaking, surnames with a nasal in E1 undergo rendaku moderately due to a lack of Identity Avoidance effects, which would otherwise promote voicing. E1-nasals do trigger rendaku in surnames with certain E2 such as /ta/; I have proposed that the constraint *NVT, an extended version of the stem-internal constraint *NT, is operative and causes voicing particularly in those names. I have also discussed the difference between E1-/m/ and E1-/n/; /m/ triggers rendaku less due to an indirect influence of the historical fact that some instances of /m/ were derived from the voiced obstruent /b/.

3.3.6 E1-approximants /w/ and /y/: Neutral consonants

Based on an observation of Sugito's (1965) data, Kubozono (2005) argues that the approximants /r/, /w/ and /y/ in E1 pattern with voiced obstruents in that they inhibit the application of rendaku. The patterns are illustrated in (13), which shows the numbers of surnames with an approximant in

[24]The experimental results show that there is no significant rendaku dampening effect of E1-/m/. See Sections 4.2.4.2 and 4.3.5 for discussion.

E1 and /ta/ as E2 by rendaku application, repeated here as (36).

(36) No. of /ta/-surnames by E1-approximant and rendaku (Sugito 1965)

E1 cons.	**t**a	**d**a
/r/	31	3
/w/	7	2
/y/	8	0

However, one should be cautious in making a generalization about all surnames out of these data, given that each E2 morpheme may show some idiosyncrasy with respect to voicing and that the total number of surnames with E1-/w/ and E1-/y/ are especially small. In fact, the results of a corpus study and an experiment reveal that /w/ and /y/ in E1 do not exert particularly strong blocking effects on rendaku (see Sections 4.2.4.2 and 4.3.4.2). In the experiment, the average rate of rendaku application in surnames with /w/ and /y/ in E1 turns out to be slightly lower than that in surnames with an E1-nasal, which moderately voice (see Section 3.3.5 above); yet it is still higher than the rendaku rate of surnames with a voiced obstruent in E1, which has strong rendaku-blocking effects by the application of Lyman's Law. I propose that /w/ and /y/ are "neutral" consonants; they do not incur a violation of OCP(voice), the Identity Avoidance requirement, nor the extended version of Post Nasal Voicing *NVT. Surnames with those consonants in E1 simply waver in rendaku application. (As is discussed below, rendaku is still triggered by a morphophonological requirement for voicing to occur in a compound.)

3.3.7 E1-approximant /r/: Inferred phonotactics *rVD

Among the approximants, /r/ needs special attention; unlike /w/ and /y/, the liquid generally inhibits rendaku application in the corpus, as is described by Kubozono (2005) (see Sections 4.2.4.2)[25] We have seen above that rendaku is blocked when a surname has a voiced obstruent in E1, since creating another voiced obstruent within the name would violate OCP(voice), which

[25]Note, however, that the effects are not found in an experiment testing the productivity. See Section 4.3.5 for discussion on the difference between the corpus study results and the experimental results.

bans the occurrence of multiple voiced obstruents. The liquid /r/, however, should not be relevant to the OCP(voice) constraint. Why, then, does it affect the voicing alternation of an obstruent in the second element? Here again, I propose an explanation based on stem-internal phonotactics. I claim that a sequence of /r/ and a voiced obstruent within one stem is underrepresented in the lexicon due to special distributional restrictions on these consonants.

One common characteristic of voiced obstruents and /r/ in Japanese phonology is that their distribution is limited to certain prosodic contexts; generally speaking, they do not appear in word-initial position.[26] The ban on initial voiced obstruents dates back to the time of Old Japanese, in which no voiced obstruent was allowed word-initially (Hashimoto 1938/1950; Unger 1977). Although the restriction is somewhat alleviated, initial voiced obstruents are still very marked in present-day Japanese (Martin 1987:30).[27] /r/ was also banned word-initially in Old Japanese (Hashimoto 1938/1950; Frellesvig 2010), and it continues to be disallowed in the language today, with only few native words begining with the consonant. Kuginuki (1982) conducts a corpus study of Old Japanese stems containing /r/; not only does he confirm the fact that /r/ did not occur in the initial position, but he also shows that it was most likely to occur at the right edge of a stem. It is shown that, among 394 stems containing /r/, the phoneme appears in the last syllable about 85% of the time.[28] Kuginuki (1982) then claims that Proto-Japanese did not have the liquid in its original consonant inventory and later developed it as a new phoneme in order to increase the length of words, which were mostly one or two moras, by attaching syllables with onset /r/. Along the same lines, Labrune (2014) proposes that [r] emerged as an epenthetic consonant, which served as a hiatus breaker. She argues that the language was predominantly a suffixing language, and that the new sound [r] was inserted when vowel-initial suffixal elements were attached to vowel-ending

[26]Another interesting fact about /r/ regarding its similarity to voiced obstruents is that it was historically subject to Obligatory Contour Principle effects. According to Labrune (2014) (also see references therein), in Old Japanese, the occurrence of multiple /r/ within a stem was categorically banned in Old Japanese.

[27]See Tanaka and Yashima (2013) for experiments testing the psychological reality of the restriction in the grammar of present-day Japanese speakers. See Westbury and Keating (1986) for a phonetic motivation of the restiction on initial voiced obstruents in general.

[28]The percentage ranges from 84% to 88% depending on what word classes are included.

words. Their proposals explain why the liquid appears typically at the right edge; since /r/ was used primarily when a syllable was added to the end of a word, it necessarily occurs near the right edge (e.g. word formation with a hypothetical noun and a hypothetical suffixical element: /sana/ + /u/ → [sana-**r**u], by /r/-insertion).[29]

Assuming that these hypotheses are on the right track and that the general distributional patterns of voiced obstruents and /r/ still hold true in present-day Japanese, we can now infer stem-internal phonotactic restrictions on the two types of sounds. Firstly, their cooccurrence within a stem is somewhat restricted. Stems are mostly bimoraic or trimoraic in present-day Japanese, and four mora stems are usually (etymologically) compounds (see Kubozono 2015). Given that voiced obstruents and /r/ are already banned initially, they usually cannot cooccur in bimoraic stems, which may only have two consonant slots at most due to prosodic structure requirements of the language (e.g. unattested *[rada], *[gora]).[30] Secondly, given that voiced obstruents are by far the most infrequent consonants in the lexicon (Labrune 2012:100), and that /r/ predominantly occurs at the right edge of a stem (see Kuginuki 1982 for the Old Japanese data), sequences of /r/ and a voiced obstruent (occurring in this particular order) in trimoraic stems should be particularly underrepresented (e.g. unattested */taraga/). The restriction only has gradient effects and we do find a handful of noun stems that go against this generalization, as in /karada/ 'body' and /kurage/ 'jellyfish.' Other exceptions mostly come from verb stems, such as /erab-u/ 'to select' and /narab-u/ 'to line up.'[31] Most importantly, however, sequences of /r/ and a voiced obstruent,

[29]Some stems ending in [ru] in present-day Japanese could possibly be analyzed as having been derived through suffixation for the purpose of increasing word length, as is claimed by (Kuginuki 1982). /yo/ is a morpheme for 'night,' which is most often used in complex words such as /yo-naka/ 'night-center (time around/after midnight).' /yoru/, another morpheme for 'night,' is more often used by itself. Shogakukan's Japanese Dictionary (Nihon Kokugo Daijiten Dainihan Henshū Iinkai 2000) gives several hypotheses on its etymology; one of them states that /yoru/ was derived from combining /yo/ 'night' and the suffixal element /-ru/, of which the meaning is not given. Similarly, /hiru/ 'afternoon' is analyzed as being etymologically composed of /hi/ 'sun/day' and /-ru/.

[30]There exist certain words that violate the ban on initial voiced obstruents and have /r/ in the second mora, as in the noun /bara/ 'rose' and the suffixed verb stem /de-ru/ 'to exist,' which mostly emerged through some historical quirk. Mimetic words constitute a class of exceptions. To the best of my knowledge, there are no bimoraic stems, including mimetics, which have initial /r/ and a voiced obstruent in the second mora.

[31]Most of these words (both noun stems and verb stems) are believed to be etymologically compounds. See Nihon Kokugo Daijiten Dainihan Henshū Iinkai (2000).

which are both independently subject to strict distributional restrictions, are expected to be rare as compared to other consonant sequences.

On the basis of such inferred phonotactics, I propose a constraint *rVD, which prohibits a voiced obstruent (represented here with a capital "D") preceded by a syllable with onset /r/ within a stem.[32] For a compound surname, the stem-internal constraint applies to the whole compound, and rendaku is blocked when E1's last syllable is /r/, in order to prevent a would-be [r...D] sequence, as shown in (37).

(37) Rendaku inhibited in surnames with /r/ in E1 due to *rVD

a.	hi̲ra	+	ta	→	hi̲ra-ta	*hi̲ra-**d**a	'flat-paddy'
b.	hi̲ro	+	kawa	→	hi̲ro-kawa	*hi̲ro-**g**awa	'large-river'

It is important to note that *rVD penalizes sequences of a liquid and a voiced obstruent occurring specifically in the order of /r/ and /D/, and not if the order is reversed (i.e., [DVr]). As is pointed out by Kubozono (2005), /r/ is particularly interesting in that it inihibits rendaku application when appearing in E1 of a surname, as shown above, but it does not act as such a strong rendaku blocking factor when appearing in E2 of a surname, as shown in (38).

(38) Rendaku not inhibited if /r/ appears in E2 of a surname

a.	matu	+	ha̲ra	→	matu-**b**ara	*matu-hara	'pine-field'
b.	ko	+	ho̲ri	→	ko-**b**ori	*ko-hori	'small-moat'

This suggests that the cooccurrence of /r/ and a voiced obstruent within a stem is not strictly banned per se. As stated above, as far as two mora stems are concerned, the two types of sounds do not usually occur side by side in a stem. From those data, Japanese speakers may still come up with a phonotactic generalization that their coexistence is relatively infrequent, which can be represented by a constraint like *$[r...D]_{stem}$/$[D...r]_{stem}$. However, giving high weight to such a constraint in the grammar would wrongly predict that rendaku will be blocked in surnames with

[32]However, see Sections 4.2.4.2 and 4.3.5 for discussion on whether the constraint is psychologically real in the grammar of present-day Japanese speakers.

/r/ in E2 as in the examples in (38).[33] The constaint should thus be relatively low ranked, and not interfere with rendaku application.[34]

To summarize, the rendaku patterns of surnames with /r/ in E1 can also be accounted for by stem-internal phonotactics. I have claimed that stems containing the liquid /r/ and a voiced obstruent in this particular order are especially underrepresented in the native lexicon because of the combination of three factors: (i) /r/ and voiced obstruents are banned initially (Hashimoto 1938/1950), (ii) /r/ is generally aligned to the right edge of a stem (Kuginuki 1982), and (iii) voiced obstruents are generally infrequent (Labrune 2014). The phonotactic constraint *rVD, which can be inferred from the lexicon, applies to stems, blocking rendaku application in compound surnames with /r/ in E1's last syllable.[35]

3.3.8 Why rendaku in a surname?: [+voice]$_{\mathfrak{R}}$ and REALIZEMORPHEME

I have claimed that compound surnames are treated as single stems and that rendaku in surnames is driven by stem-internal phonotactics. One may wonder, however, why rendaku applies at all in a surname, if the voicing alternation is a morphophonological operation in compound formation and if the surname is regarded as a stem. It should be made clear that, as shown in (27), compound surnames are recursive stems; they are labeled as a stem, but they have a compound structure, being composed of stems. This means that, just like a regular compound word, a compound surname is derived through morphological compound formation, which requires voicing of

[33]Since the constraint bans the cooccurrence of /r/ and a voiced obstruent within a stem, its high ranking would predict that rendaku will be blocked even in a regular compound with E2 containing /r/. There exist, however, many compounds violating the constraint, especially those with a verb or a deverbal noun as E2: e.g. /oya-tori/ → [oya-**d**ori] 'parent-bird,' /ki-hori/ → [ki-**b**ori] 'wood-carving,' /nasi-kari/ → [nasi-**g**ari] 'pear-picking.' (Verb stems often contain /r/; Labrune 2014.)

[34]Interestingly, /karada/ 'body,' which is given above as one of the exceptional stem nouns violating *rVD, undergoes a metathesis to become [kadara] in a number of dialcets that are not closely related (see Nihon Kokugo Daijiten Dainihan Henshū Iinkai 2000). This may be taken as evidence for the dispreference for [r...D] over [D...r] sequences.

[35]The validity of the claim about the distributions of /r/ and voiced obstruents in present-day Japanese made based on Old-Japanese patterns remains to be tested. It would require an extensive study of native Japanese stems, which I leave for future research.

E2's initial consonant, namely, rendaku application. In other words, rendaku in surnames is independently motivated as a morphophonological process, and stem-internal phonotactics and other phonological requirements only come into play as promoting or inhibiting factors in its realization.

As a formal account of the voicing phenomenon in regular compounds, Ito and Mester (1986, 1998, 2003) propose that rendaku is a realization of a feature-sized linking morpheme $[+\text{voice}]_{\mathfrak{R}}$, which functions to connect the components of a compound.[36] This $[+\text{voice}]_{\mathfrak{R}}$, which is introduced in the input by morphological compounding, is required to be realized in the output by the constraint REALIZE-MORPHEME, defined in (39) below.

(39) REALIZE-M(ORPHEME):
Every morpheme in the input has a nonnull phonological exponent in the output (Ito and Mester 2003)

The constraint is satisfied if the linking morpheme's [+voice] feature specification is realized in an output segment, as in the rendaku form of a compound: e.g. /maki+$[+\text{voice}]_{\mathfrak{R}}$+susi/ → [maki-**z**usi] 'rolled-sushi.'[37]

Following Ito and Mester (1986, 1998, 2003), I propose that compound surnames also come with $[+\text{voice}]_{\mathfrak{R}}$ in the input as a result of compounding. As is shown in the tableau in (40), rendaku applies if REALIZE-MORPHEME outranks Ident(voice), which penalizes an output segment differing from its input correspondent in the voicing feature.

(40) Rendaku application in a surname by high-ranked REALIZE-MORPHEME

$[\ [\text{sawa}]_{\text{stem}} + [+\text{voi}]_{\mathfrak{R}} + [\text{ta}]_{\text{stem}}\ ]_{\text{stem}}$	REALIZE-M	IDENT(voice)
(a) sawa-ta	*!	
☞ (b) sawa-da		*

[36] See the historical development of rendaku discussed in Section 2.6.1 for an explanation of how such a morpheme emerged in the Japanese language. For similar linking morphemes in other languages, see Ito and Mester (2003:83-85) and Labrune (2016) among others.

[37] $[+\text{voice}]_{\mathfrak{R}}$ is placed between the first element and the second element of a compound in the input. I assume that LINEARITY (McCarthy and Prince 1995) is ranked high in the grammar, which ensures that the voicing feature can only be realized on the initial obstruent of E2.

As discussed above, whether voicing is realized or not in each surname is also affected by many other phonological and lexical factors. In Chapter 5, I will give a detailed account of how these factors interact with each other, producing the variable rendaku patterns.

3.3.9 Summary

Based on the claim that surnames are stems with a compound structure, I have proposed that their rendaku patterns can be attributed to the application of stem-phonology. I have shown that the proposal answers the questions raised earlier about why the voicing alternation shows some peculiarities in surnames that are not seen in regular compounds. The proposal also makes some new predictions about how rendaku applies in surnames that are not descibed in detail in previous studies. The effects of stem-internal phonological restrictions in rendaku in surnames are summarized in (41).

(41) Stem-internal phonology affecting rendaku in surnames

- OCP(voice)
  - ▹ Rendaku is blocked by voiced obstruents in E1 (Strong Lyman's Law)
- Identity Avoidance
  - ▹ Rendaku is triggered by homorganic voiceless obstruent sequences such as /s...s/, /t...t/, /k...k/ and /s...t/ occurring across E1 and E2
  - ▹ Rendaku is not triggered by non-homorganic obstruent sequences such as /k...t/ and /t...k/ in E1 and E2 (/k/'s peculiarity solved)
  - ▹ Rendaku is not triggered by sonorant-obstruent sequences such as /y...t/ and /w...t/
- *NVT and *mVD
  - ▹ Rendaku is triggered by nasal-obstruent sequences such as /m...t/ and /n...t/ in certain surnames
  - ▹ /m...t/ triggers rendaku less than /n...t/
- *rVD
  - ▹ Rendaku is blocked by liquid-obstruent sequences such as /r...t/

3.4 Analysis: Deriving prosodic factors

In this section, I will argue that the prosodic patterns of compound surnames also follow from the application of stem-phonology. I will first show that the predictability of accent location in compound surnames can simply be derived from the application of the antepenultimate accent rule (Martin 1952; McCawley 1968), which is the default for single stems. I will further propose a foot-based account of the correlation between rendaku application and accentedness. I will show that Ito and Mester's (2016) analysis of (un)accentedness in Japanese single stems can be directly extended to compound surnames, deriving the accent-rendaku correlation.

3.4.1 Predictability of accent location: Default antepenultimate

As we have seen in Section 2.4.1, accentuation in compound surnames is fairly predictable; they are either unaccented or have accent on the syllable containing the antepenultimate mora. This forms a contrast with other uninflected nouns, which show variability in accent location. The predictability of the place of accent highlights another peculiarity of compound surnames; they do not follow the usual compound accent rule. Descriptively speaking, for regular compounds with a short E2 morpheme (one or two moras), the compound accent falls at the edge of the boundary between the components, namely on the final syllable of E1 or the initial syllable of E2 (see Kubozono 2008 and Kawahara 2015b for an overview of studies on accent in Japanese). Compound surnames, however, receive antepenultimate accent, often going against the expected patterns, as shown in (17), repeated here as (42).

(42) No compound accent in compound surnames

a.	síro	+	tá	→	^{3}**sí**ro-ta	^{*2}si**ró**-ta		'white-paddy (surname)'
b.	náka	+	mori	→	^{3}na**ká**-mori			'center-forest (surname)'
c.	miya	+	máe	→	^{3}mi**yá**-mae	^{*2}miya-**máe**		'shrine-front (surname)'

Except for (42b) where the compound accent rule would place accent on the antepenultimate syllable anyway, the location of accent in a surname differs from what would be expected in a

compound. In (42a), accent falls on the syllable containing the antepenultimate mora of the entire name, instead of the final syllable of E1. In (42c), the E2 [2]máe] 'front' is a type of morpheme which usually preserves its initial accent when appearing in a compound (unless the compound is unaccented), yet again accent falls on the antepenultimate syllable of the entire name.

This seemingly peculiar accent pattern of surnames is naturally derived once we assume that they are treated as single stems in phonology; compound surnames simply follow the antepenultimate accent rule (McCawley 1968), which is the most common and productive accent assignment in Japanese stems (Kubozono 2008). We have seen earlier (Section 2.4.1) that the location of accent in uninflected nouns excluding proper names is unpredictable, as it may fall on any syllable of the word. It is not the case, however, that all patterns are equally attested. Kubozono (2006, 2008) shows that, among accented trimoraic stems, the most common accent position is antepenult both in the native (59%) and foreign (96%) strata.[38] These facts suggest that antepenultimacy is the "default" for Japanese stems. A simple statement that accent falls on the syllable containing the antepenultimate mora (and on the initial syllable for words shorter than three moras) can account for the patterns of many accented native stems. The generalization holds especially true in loanwords, which arguably lack accent specifications in the input (see McCawley 1968; Shinohara 2000; Kubozono 2006). Turning back to compound surnames, I argue that their accent patterns can be explained in terms of the assignment of stem accent. A surname is formed by compound formation, and the word itself does not have any underlying accent specification. Regular compounds would receive accent by the compound accent rule; but compound surnames, which are labeled as stems, receive antepenultimate accent, just like other regular stems.

In the sections below, I will propose a constraint-based analysis of accent in surnames. I will show that the analysis can not only capture their basic antepenultimate accent pattern but also predict the unaccented pattern in names with rendaku, solving the problem of the correlation between accent and voicing.

[38] About 59% of trimoraic native stems show antepenultimate accent, 33% show penultimate accent and 9% show final accent.

3.4.2 Antepenultimacy and unaccentedness in stems: Ito and Mester (2016)

I have claimed that surnames are represeted as stems in the grammar and thus follow the accent patterns of stems. In this subsection, I will first review Ito and Mester's (2016) formal account of the accent patterns of Japanese stems, specifically antepenultimacy and unaccentedness, based on OT constraints.

Ito and Mester (2016) start with the observation that phonological length (or mora counts) has an effect on accentedness in Japanese stems, especially in loan items; four mora words are disproportionately unaccented while others are mostly accented.[39] They claim that the length-based patterns are not accidental but structural in origin. To go straight to the bottom line, antepenultimacy is derived from trochaic footing with final extrametricality as in [(μ́μ)μ], whereas unaccentedness arises from exhaustive footing of a word and a failure in placing accent on either one of two consecutive feet: e.g. [(μμ)(μμ)]; cf. *[(μ́μ)(μμ)], *[(μμ)(μ́μ)]. Under this analysis, unaccented words have a specific foot structure (rather than being unfooted; [μμμμ]) but receive no accent as a result of constraint interactions. The problem with having accent with [(μμ)(μμ)] is that one constraint prohibits accent on the first foot, and another constraint prohibits accent on the second foot. When a constraint which requires a word to have accent (dubbed WORDACCENT; see the detailed definition below) is outranked by these constraints, the word is rendered unaccented. The analysis answers the question of why four mora words, which can easily be footed exhaustively while respecting foot-binarity as in [(μμ)(μμ)], are predominantly unaccented.

Let us take a look at how their stem accent grammar generates unaccentedness as well as antepenultimacy, by comparing accent assignment in three mora and four mora loan items. The definitions of the key constraints are given in (43).

[39] The accentedness tendencies differ depending on vocabulary strata. In the native stratum, there are also not a few three mora items that are unaccented. See Ito and Mester (2016:513–514) for a suggestion about the historical origin of such patterns.

(43) Key constraints in Ito and Mester's (2016) stem accent grammar

a. WORDACCENT (WDACC):
A prosodic word contains a prominence peak. Violated by prosodic words not having a prominence peak (peak = primary stress or pitch accent, in Japanese: High*Low)

b. RIGHTMOST:
*Ft'...Ft...]$_{\omega}$. Violated by any foot following the head foot within the prosodic word (ω) (McCarthy 2003:111).

c. NONFINALITY(FT') (NONFIN(FT')):
*Ft']$_{\omega}$. Violated by any head foot that is final in its prosodic word (ω) (Prince and Smolensky 1993/2004:45).

d. INITIALFOOT (INIFT):
*$_{\omega}$[o. A prosodic word begins with a foot (Ito and Mester 1992/2003:31; McCarthy and Prince 1993:81). Violated by any prosodic word (ω) whose left edge is aligned with the left edge, not of a foot, but of an unfooted syllable.

e. PARSE-σ:
*o. All syllables are parsed into feet (Prince and Smolensky 1993/2004:62). Violated by unfooted syllables.

f. NOLAPSE:
*oo. Syllables are maximally parsed. Violated by two consecutive unparsed syllables.

g. FOOTBINARITY (FTBIN):
Feet are minimally binary at some level of analysis (μ, σ). Violated by unary feet.

Also note that the constraint TROCHEE, which requires head-initiality of feet (e.g. ✓[(μ́μ)]; cf. *[(μμ́)]), is assumed to be ranked high in the grammar (crucially above IAMB), which ensures that all feet are trochaic.[40]

A grammar with the ranking shown in (44) assigns antepenultimate accent to a three mora word, as is illustrated in the tableau with the input /banana/ 'banana.'

[40]More precisely, Ito and Mester (2016) proposes MORAICTROCHEE, a mora-based version of the constraint. The distinction is not crucial for our purposes.

(44) Antepenultimate accent in a three mora stem

/banana/	NoLapse	FtBin	Nonfin(Ft')	Rightmost	IniFt	WdAcc	Parse-σ
a. [1](bana)(ná)		*!	*				
b. [3](bána)(na)		*!		*			
c. [0](bana)(na)		*!				*	
d. [2]ba(nána)			*!		*		*
e. [0](bana)na						*!	*
☞ f. [3](bána)na							*

Exhaustive footing of a three-mora word produces a unary foot, as in (a), (b) and (c); these candidates are all eliminated due to a violation of high-ranked FtBin. The optimal foot structure is an initial binary foot with final extrametricality, as in (e) and (f). Candidate (f) with penultimate accent wins over unaccented (e), which violates WdAcc.

The analysis thus captures the generalization that three mora stems mostly have penultimate accent. How about four mora words, which are often unaccented? The same grammar with the exact same ranking actually derives unaccentedness in four mora stems. This is illustrated in tableau (45) with the input /amerika/ 'America.'

(45) Unaccentedness in a four mora word

/amerika/	NoLapse	FtBin	Nonfin(Ft')	Rightmost	IniFt	WdAcc	Parse-σ
a. ^{2}(ame)(ríka)			*!				
b. ^{4}(áme)(rika)				*!			
c. ^{3}a(méri)ka					*!		**
☞ d. ^{0}(ame)(rika)						*	
e. ^{4}(áme)rika	*!						**

In the case of a four-mora word, antepenultimacy is not the optimal accent pattern; candidate (c) loses due to a violation of INITIALFOOT, which penalizes a prosodic word that begins with an unfooted syllable. Candiates with exhaustive footing, as in (a), (b) and (d), all satisfy FTBIN. However, candidate (a) with penultimate accent violates NONFINALITY(FT'), which penalizes a final head foot. Candidate (b) with initial accent also violates RIGHTMOST, which requires a head foot to be the rightmost foot of the word. Since NONFINALITY(FT') and RIGHTMOST both outrank WORDACCENT, (d) with the same foot structure but with no accent is the optimal candidate. In other words, when a four mora word is exhaustively footed with two consecutive bimoraic feet, there is a tension between an initial head foot ([(μ́μ)(μμ)]) and a final head foot ([(μμ)(μ́μ)]). Since either option would incur a violation of high-ranked NONFINALITY(FT') or RIGHTMOST, the conflict is resolved by rendering the word unaccented ([(μμ)(μμ)]) at the cost of violating lower-ranked WORDACCENT.

The analysis thus nicely captures the length-based tendencies of stem accent: antepenultimacy in three mora words and unaccentedness in four mora words.[41] Ito and Mester (2016) further show that minimal reranking of some of the constraints can even capture variation, deriving other

[41]Monomoraic and bimoraic words are also mostly accented, especially in loans (e.g. ^{0}[dó] 'do (musical note),' ^{0}[mémo] 'memo'), although the analysis shown here predicts that they will be unaccented due to the ranking NONFIN(FT') ≫ WDACC. Ito and Mester (2016) proposes a high-ranked constraint MINIMALWORDACCENT, which requires that prosodically minimal (monomoraic and bimoraic) words be accented. See Ito and Mester (2016:489–491) for discussion on the motivation for the constraint.

patterns that are somewhat rarer. For example, as lexical variation, four mora words may also show antepenultimate accent, as in ^{3}[papúrika] 'paprika.' Ranking INIFT below WDACC predicts this antepenultimate pattern in a four mora word, as shown in (46). (Note that the ranking RIGHTMOST, NONFINALITY(FT') ≫ WORDACCENT is kept as is.)

(46) Antepenultimate accent in a four mora word: WDACC≫INIFT

/papurika/	NOLAPSE	FTBIN	NONFIN(FT')	RIGHTMOST	WDACC	INIFT	PARSE-σ
a. ^{2}(papu)(ríka)			*!				
b. ^{4}(pápu)(rika)				*!			
☞ c. ^{3}pa(púri)ka						*	**
d. ^{0}(papu)(rika)					*!		
e. ^{4}(pápu)rika	*!						**

Now that a violation of INIFT is not crucial, it is better to have accent with a medial trochaic foot and a final extrametrical syllable, as in (c). Note that this ranking still produces antepenultimacy in three mora words since reranking INIFT and WDACC does not affect the selection of a winner, as can be seen in the violation profiles of the candiates in (49) above.

It is worth noting here that the analysis shown thus far is about simplex stems. Morphologically complex words may show yet different patterns. Ito and Mester (2016) observe that truncated compounds, which are three or four moras long, are always unaccented regardless of their prosodic length. As is shown in (47), three-mora items as well as four-mora items are unaccented.

(47) Unaccentedness in truncated compounds

a.	famirii-máato	→	^{0}fami-ma	'Family Mart'
b.	sukuriin-syótto	→	^{0}suku-syo	'screenshot'
c.	hebii-métaru	→	^{0}hebi-meta	'heavy metal'
d.	rabu-kómedii	→	^{0}rabu-kome	'love comedy'

Ito and Mester (2016) argue that the unaccented patterns of truncated compounds are also rooted in exhaustive footing. With independent evidence reported by (Poser 1984), they claim that there is a requirement for each element of a compound to be parsed into a foot, which is stated in the form of the constraint in (48).

(48) LEXICALFOOT (LEXFT):
Every lexical morpheme (i.e., full content morpheme, not grammatical formative) minimally projects its own foot.

If LEXFT is ranked high in the grammar (crucially above FTBIN, and above WDACC by transitivity), each element must be independently footed. This results in exhaustive footing of a word, and unaccentedness follows even in a trimoraic item, as is illustrated in the tableau in (49). (Note that WDACC is ranked low, crucially below NONFIN(FT') and RIGHTMOST in this case, which derives unaccentedness. Its ranking with respect to INIFT is not crucial here.)

(49) Unaccentedness in a three mora truncated compound due to top-ranked LEXFT

/fami-ma/	LEXFT	NOLAPSE	FTBIN	NONFIN(FT')	RIGHTMOST	INIFT	WDACC	PARSE-σ
a. ^{1}(fami)-(má)			*	*!				
b. ^{3}(fámi)-(ma)			*		*!			
☞ c. ^{0}(fami)-(ma)			*				*	
d. ^{2}fa(mí-ma)	*!*			*		*		*
e. ^{0}(fami)-ma	*!						*	*
f. ^{3}(fámi)-ma	*!							*

As is briefly discussed by Ito and Mester (2016:511), this categorical unaccentedness in truncated compounds raises another interesting question: why does their accentuation pattern differ from that of regular (non-truncated) compounds? Recall that regular compounds of the same configuration (i.e. 2μ + 1μ or 2μ + 2μ) may be unaccented, but they may also be accented, with accent falling either on the last syllable of the first element or on the initial syllable of the second ele-

ment (see (16) and (17) in Section 2.4.1; also see (42) in Section 3.4.1). Most previous analyses of accent in regular compounds also assume an inviolable requirement similar to LEXFT and basically the same foot structures for three and four mora regular compounds (see e.g. Kubozono 1995, 1997, 2008; Tanaka, S. 2001; cf. Alderete 2015 for a slightly different footing analysis). The fact that they are not always unaccented even with exhaustive footing suggests that there are additional factors (e.g. some high-ranked constraints or accent assignment through a different subgrammar) which make them accented, with a specific accent location.[42] On the other hand, truncated compounds are still somewhat like simplex stems, in the sense that they are not subject to such additional accenting requirements. Giving a full account of accentuation in regular compounds, or explaining why truncated compounds behave differently from regular compounds is beyond the scope of this dissertation. Here, I simply make the following assumption. As is suggested by previous studies, each element of a compound must be parsed into a foot (i.e., LEXFT is ranked high); but regular compounds are subject to some additional accentuation requirement, which can overturn the anti-accenting effect caused by exhaustive footing.

To summarize, the grammar proposed by Ito and Mester (2016) derives the two basic accent patterns in Japanese stems, namely, antepenultimacy and unaccentedness, through interactions of prosodic constraints. In terms of metrical feet, antepenultimate accent is derived when a word has one foot with extrametricality, whereas unaccentedness arises when a word is exhaustively parsed into feet. The generalizations do not apply to regular compounds, which show different accentuation patterns. Below, I will extend Ito and Mester's (2016) analysis to the case of compound surnames, treating them as stems.

[42] Alderete (2015) proposes a high-ranked constraint PRWDHEADACCENT which requires the head of a prosodic word compound to have a peak prominence. Assuming that Japanese compounds are right-headed, he explains why compound accent falls on the initial syllable of E2, as in ^{4}[minami-**á**merika] 'South America,' derived from unaccented ^{0}[amerika]. Under the analysis, the head, or E2, is required to receive accent due to high-ranked PRWDHDACC. Alderete further argues that, when accent falls on the last syllable of E1, as in ^{3}[kabu**tó**-musi] 'armor-insect (beetle),' there is a mismatch between the actual morpheme boundary and the prosodic structure, and the accented syllable is prosodically part of the second element, satisfying PRWDHDACC. The assumed prosodic structure is, e.g., ^{3}[{(kabu)}$_{E1}${(tó-mu)si}$_{E2}$]. Although it has not been tested against all kinds of compounds, the merits of the proposal are: (i) it is compatible with Ito and Mester (2016) and can analyze simplex stems and regular compounds with one and the same grammar, and (ii) there is no need to resort to trochaic-iambic foot flipping, assumed by most previous studies (Kubozono 1995, 1997; Tanaka, S. 2001): e.g. ^{3}[ka(butó)-(musi)].

3.4.3 Stem grammar applied to compound surnames: Default antepenultimacy

As stated above, Japanese compound surnames are either unaccented or have antepenultimate accent. Based on the proposal that compound surnames show the accent patterns of stems, I argue that their antepenultimacy and unaccentedness are derived from the same principle as the one shown by Ito and Mester (2016) in their account of Japanese stems. More precisely, I claim that compound surnames simply go through Ito and Mester's (2016) stem accent grammar, just like regular stems.

In deriving the accent patterns of compound surnames through Ito and Mester's (2016) grammar of stem accent, I first make two assumptions. The first is about which subgrammar is to be adopted. In order to capture variation in their data, Ito and Mester (2016) propose four subgrammars with slightly different constraint rankings. I have introduced two with minimal reranking of WORDACCENT and INITIALFOOT. As is shown above, a grammar with the ranking INIFT ≫ WDACC generates antepenultimacy in three mora stems and unaccentedness in four mora stems (e.g. ^{3}[(bána)na], ^{0}[(ame)(rika)]), while a grammar with the other ranking WDACC ≫ INIFT predicts that both types will have antepenultimate accent (e.g. ^{3}[(bána)na], ^{3}[pa(púri)ka]). Ito and Mester (2016) assume the former to be the most basic grammar, calling it "the default unaccented system," in light of the fact that unaccentedness is found in the majority of four mora items (about 52.6% of four mora words with light syllables, or LLLL, in their corpus; n = 228). Compound surnames, however, show different tendencies (despite having the same accentuation systems); the majority of four mora surnames are accented. In Zamma's (2005) data of four mora surnames with E2-/kawa/ shown in (20) and (53), 20 have antepenultimate accent, whereas 12 are unaccented (ignoring variable ones). (The tendency for antepenultimacy looks even more prominent once we consider the fact that unaccentedness in surnames is caused by rendaku; see below for the details.) Similar data are found in the accent patterns of given names. Sugawara (2012) shows that about 75% of four mora given names (n = 160) have antepenultimate accent, and even more strikingly, 100% of those with only light syllables (LLLL) (n = 83) show antepenultimacy. Based on these observations that names tend to be generally more accented than regular stems, I claim that the

default ranking for compound surnames (and proper nouns in general) is WDACC ≫ INIFT.[43,44]

The second assumption is about the applicability of the constraint LEXFT. As discussed above, most previous studies including Ito and Mester (2016) posit that each element of a compound necessarily has its own foot. I claim that this foot structure requirement on compounds does not apply to compound surnames for two reasons. The first is the data patterns themselves. As I have shown in (42), compound surnames do not follow the usual compound accentuation pattern, which is generally aligned to the boundary between the elements. The fact that accent falls on the antepenultimate syllable, ignoring the presence of a morpheme boundary, suggests that footing in compound surnames is not constrained by internal morphological structures. The second is the semantic properties of the elements of compound surnames. As discussed in Section 3.2.1, compound surnames are not semantically compositional. Each element of a compound can have a meaning of its own when appearing as a regular word by itself, as in /yama/ 'mountain' and /kawa/ 'river.' However, a compound surname formed with those elements, as in /yama-kawa/ 'mountain-river (surname),' simply refers to an individual with that name, and never has a compositional meaning that a regular compound composed of the same morphemes, e.g. /yama-kawa/ 'mountain river' or 'mountains and rivers,' would convey (unless one is talking about the etymological origin of the name). To put it differently, morphemes do not have meanings of their own when they appear in a compound surname.[45] Ito and Mester (2016) define LEXFT as a constraint penalizing

[43]Another possible way to capture the accenting tendency in proper nouns is to posit a name-specific constraint NAMEACCENT, which requires that proper nouns have a prominence peak, and rank the constraint above INIFT, while keeping the default INIFT ≫ WDACC ranking. For our purposes, the two analyses are notational variants, and for the sake of simplicity, I adopt the WDACC ≫ INIFT ranking, which is one of the original rankings proposed by Ito and Mester (2016), instead of introducing a new constraint.

[44]Ito and Mester (2016) have two subgrammars where WDACC outranks INIFT, which they call the "antepenulimate system" and the "weak antepenultimate system." The two grammars may produce different outcomes in four mora items with different syllable weights. However, my analysis shown below is only concerned with names with light syllables (LLL and LLLL) and both grammars make the same predictions. For the sake of simplicity, I adopt a ranking corresponding to their weak antepenultimate system, which only involves reranking of WDACC and INIFT from the default unaccented system.

[45]One possible interpretation of the semantic non-compositionality of compound surnames is that content morphemes lose their meanings as a result of name formation. Another interpretation is that morphemes which appear in names exist independently in the lexicon as some kind of formatives specifically for surnames without full content meanings.

a full content morpheme that is not projecting its own foot, as shown in (48) above. I argue that elements of compound surnames do not count as "full content morphemes" and are exempt from the requirements enforced by LEXFT.[46]

With these assumptions in mind, let us see how Ito and Mester's (2016) stem grammar assigns accent to compound surnames. The tableau in (50) shows that it derives antepenultimate accent for a three mora compound surname such as /saka-ta/ 'slope-paddy.' (Note that rendaku is not realized in the surname as a result of constraint interactions not shown here. Here, I omit the linking morpheme [+voice]$_{\mathfrak{R}}$ in the input for the sake of simplicity.)

(50) Antepenultimate accent in a three mora compound surname

/saka-ta/	LEXFT	NOLAPSE	FTBIN	NONFIN(FT')	RIGHTMOST	WDACC	INIFT	PARSE-σ
a. ^{1}(saka)-(tá)			*!	*				
b. ^{3}(sáka)-(ta)			*!		*			
c. ^{0}(saka)-(ta)			*!			*		
d. ^{2}sa(ká-ta)				*!			*	*
e. ^{0}(saka)-ta						*!		*
☞ f. ^{3}(sáka)-ta								*

The pattern is analogous to the case of a three mora single stem like ^{3}[(bána)na] shown in (44). Due to high-ranked FTBIN, exhaustive footing of a three mora name is not viable. The optimal prosody is a trochaic foot with an extrametrical syllable, which derives antepenultimate accent, as in candidate (f). Crucially, LEXFT does not affect surnames, allowing the second element /ta/ to be unparsed.

What about four mora surnames? As is shown in (51), the grammar also assigns antepenultimate accent to a four mora surname such as /yama-kawa/ 'mountain-river.'

[46] Another way of expressing the inapplicability of LEXFT is that the constraint can only look at the outermost label of recursive stems: [[A]$_{\text{stem}}$ -[B]$_{\text{stem}}$]$_{\textbf{stem}}$.

(51) Antepenultimate accent in a four mora compound surname

/yama-kawa/	LexFt	NoLapse	FtBin	Nonfin(Ft')	Rightmost	WdAcc	IniFt	Parse-σ
a. ^{2}(yama)-(káwa)				*!				
b. ^{4}(yáma)-(kawa)					*!			
c. ^{0}(yama)-(kawa)						*!		
☞ d. ^{3}ya(má-ka)wa							*	**
f. ^{4}(yáma)-kawa		*!						**

Notice that the pattern is similar to a four-mora single stem with antepenultimate accent, as in ^{3}[pa(púri)ka] shown in (46) above. The ranking WdAcc ≫ IniFt ensures that the word is accented, at the cost of having an unparsed syllable at the beginning. Again, the inapplicability of high-ranked LexFt to compound surnames is crucial since it allows a foot to span a boundary as in the winning candidate (d) ^{3}[ya(má-ka)wa].

We have seen that Ito and Mester's (2016) stem accent grammar with the ranking WdAcc ≫ IniFt derives antepenultimate accent in three and four mora names. I claim that antepenultimacy is the default for compound surnames, given that it is statistically the most common pattern (see above). The next question to be addressed is how unaccentedness, the other common pattern, arises in surnames. I will argue below that it is the application of rendaku which forces names to have a compound-like foot structure, which in turn causes them to be unaccented. I will show below that, with the addition of only one constraint on prosodic requirements for rendaku application, the grammar can derive unaccentedness in surnames with rendaku voicing, solving the issue of the correlation between accentedness and rendaku.

3.4.4 Deriving the rendaku-accent correlation

As discussed in Section 2.4.2, previous studies reveal an inverse correlation between rendaku application and accentedness in compound surnames (see Sugito 1965; Zamma 2001, 2005; Ohta 2013

among others). Although the degree of correlation varies greatly depending on the E2 morpheme (Zamma 2005), the general pattern is that surnames with rendaku are unaccented while those without rendaku are accented (with penultimate accent). The tables in (52) and (53) (repeated from (19) and (20) respectively) show the numbers of surnames with /ta/ and /kawa/ as E2 sorted by rendaku application and accentedness, taken from Sugito (1965) and Zamma (2005).

(52) No. of /ta/-surnames by rendaku and accentedness (Sugito 1965)

	Accented	Unaccented	Variable accent
No rendaku	**94**	13	10
Rendaku	64	**95**	56
Variable rendaku	8	0	22

(53) No. of /kawa/-surnames by rendaku and accentedness (Zamma 2005)

	Accented	Unaccented	Variable accent
No rendaku	**19**	2	0
Rendaku	1	**10**	0
Variable rendaku	0	2	3

In order to account for the patterns, one needs to understand the cause of the correlation. The question I raised earlier is whether it is rendaku application which affects accentedness or it is the presence/absence of accent which affects voicing. A clue to the question lies in the fact that antepenultimacy is the default accent pattern in surnames. Generally speaking, surnames are more commonly accented than unaccented. I have also shown that the accent grammar of Japanese stems (Ito and Mester 2016), when applied to compound surnames, derives antepenultimate accent. If antepenultimacy is the most common pattern and it comes out naturally as a consequence of the application of the stem grammar to surnames, the question now comes down to why names tend to have the more marked pattern, that is, be unaccented, particularly when rendaku voicing is present.[47]

[47] Also note that given names, which show similar accentuation as compound surnames but are not relevant to rendaku application, are predominantly accented (Sugawara 2012). See Section 3.4.3 above.

As an answer to the question, I argue that it is rendaku application which derives unaccentedness through a requirement on prosody. More specifically, I propose that the realization of rendaku, or compound voicing, demands each element of a compound surname to be parsed into a foot, just like in other types of compounds. This results in exhaustive footing of a three or four mora surname and hence unaccentedness arises. The proposal builds on the insight of Ito and Mester (2016) that a compound(-like) foot structure causes stems to be unaccented. We have seen that truncated compounds are all unaccented regardless of phonological length due to footing of each element: e.g. ^{0}[(fami)-(ma)], ^{0}[(hebi)-(meta)]. Ito and Mester (2016) further point out that many of the trimoraic native items which are unexpectedly unaccented are etymologically compounds, and suggest that the original exhaustive footing as a compound has been lexicalized.[48] Along the same lines, I claim that surnames with rendaku have a compound-like foot structure in that each element projects its own foot: e.g. ^{0}[(yama)-(**d**a)], ^{0}[(miya)-(**g**awa)]; cf. ^{3}[(sáka)-ta], ^{3}[ya(má-ka)wa]. As stated above, I follow the assumption that rendaku is a realization of a linking morpheme, which in principle links two stems to form a compound. It seems reasonable that the morpheme requires the linked elements to be prosodically full-fledged with their independent feet. As a formalization of such a prosodic requirement for rendaku application, I propose the constraint in (54).

(54) LINKFOOTEDSTEMS (LINKFTSTEMS):
If there is a phonological exponent of a linking morpheme,
it must occur between stems projecting their own feet.

The constraint is violated by a compound with a realization of a linking morpheme (e.g. rendaku or [+voice]$_{\mathfrak{R}}$) if none of the elements is footed (e.g. *[yama-**d**a]), only one of the element is footed (e.g. *[(yama)-**d**a]), or the two elements share a foot without having their own feet (e.g. *[ya(na-**g**a)wa]). Surnames without rendaku application, on the other hand, do not violate this constraint regardless of the foot structure (e.g. ✓[(sáka)-ta]); ✓[ya(má-ka)wa]). As will be

[48]The examples given in Ito and Mester (2016) are: ^{0}[(saka)(na)] 'fish' < [(sake)-(na)] 'sake-food (food that goes well with sake)'; ^{0}[(ne)(zumi)] 'rat' < [(ne)-(zumi)] < [(ne)-(sumi)] 'root-live (those that live in dark/low places).' We indeed find many other examples: e.g. ^{0}[(ma)(koto)] 'truth' < [(ma)-(koto)] 'true-thing'; ^{0}[(to)(kage)] 'lizard' < [(to)-(kage)] 'door-shadow/behind (those that live behind doors/furniture).' Some of them could still be morphologically analyzed by present-day Japanese speakers, but most of them are seen as simplex words. Ito and Mester (2016) suggest that their foot structures have become lexicalized, instead of treating them as actual compounds.

shown below, ranking this constraint relatively high in the grammar ensures that each element of a compound surname is parsed into a foot.

Before turning to the analysis of unaccentedness in surnames with rendaku, I introduce two other constraints, which we have seen already: RENDAKU, which requires the initial consonant of the second element of a compound to be voiced, and IDENT(voice), which penalizes an output segment that differs from its input correspondent in the voicing feature. RENDAKU is not a real OT constraint; rather, it is shorthand for all the factors or constraints on rendaku, including those that trigger it (e.g. REALIZE-MORPHEME and Identity Avoidance), those that inhibit it (e.g. OCP(voice)), and lexical propensities for the alternation. Since each surname may have different violation profiles for those bundled constraints, where it stands in the grammar can vary for each name. Its ranking with respect to IDENT(voice) determines whether the surname undergoes rendaku.

Let us now see how unaccentedness can be derived for a three mora compound surname with rendaku. The tableau in (55) illustrates the grammar generating unaccented ^{0}[yama-**d**a] 'mountain-paddy.' LINKFTSTEMS is ranked above FTBIN. For this particular surname, RENDAKU outranks IDENT(voice), causing the voicing alternation to be realized. For illustrative purposes, these constraints are placed at the very top in the tableau (without conclusive ranking arguments). For concreteness, the linking morpheme [+voice]$_{\mathfrak{R}}$, abbreviated as $\mathfrak{R}$, is shown in the input. LEXFT is omitted from the following tableaux, as it is not relevant to compound surnames.

(55) Unaccentedness in a 3μ surname with rendaku

/yama-ℜ-ta/	RENDAKU	IDENT(voice)	NOLAPSE	LINKFTSTEMS	FTBIN	NONFIN(FT')	RIGHTMOST	WDACC	INIFT	PARSE-σ
a. [3](yáma)-ta	*!									*
b. [0](yama)-(ta)	*!				*			*		
c. [1](yama)-(dá)		*			*	*!				
d. [3](yáma)-(da)		*			*		*!			
☞ e. [0](yama)-(da)		*			*			*		
f. [2]ya(má-da)		*		*!		*			*	*
g. [0](yama)-da		*		*!				*		*
h. [3](yáma)-da		*		*!						*

Due to RENDAKU ≫ IDENT(voice), any candidate without rendaku voicing are eliminated, as is illustrated with (a) and (b). High-ranked LINKFTSTEMS also eliminates candidates which realize rendaku without each element being parsed into a foot, as in (f), (g) and (h). Once there are two feet in a row, neither foot can hold accent and the optimal solution is to render the word unaccented, as in (e), for reasons which are familiar to us: NONFIN(FT'), RIGHTMOST ≫ WDACC.

The same analysis applies to four mora surnames with rendaku. If rendaku is realized, the word is unaccented, as shown in the tableau in (56) which selects [0][(yana)-(**g**awa)] as the winner for the input /yana-ℜ-kawa/ 'willow-river.'

(56) Unaccentedness in a 4μ mora surname with rendaku

/yana-ℜ-kawa/	RENDAKU	IDENT(voice)	NOLAPSE	LINKFTSTEMS	FTBIN	NONFIN(FT')	RIGHTMOST	WDACC	INIFT	PARSE-σ
a. ^{3}ya(ná-ka)wa	*!								*	**
b. ^{0}(yana)-(kawa)	*!							*		
c. ^{2}(yana)-(gáwa)		*				*!				
d. ^{4}(yána)-(gawa)		*					*!			
☞ e. ^{0}(yana)-(gawa)		*						*		
f. ^{3}ya(ná-ga)wa		*		*!					*	**
g. ^{4}(yána)-gawa		*	*!	*						**

Rendaku realization with footing across a morpheme boundary as in (f) ^{0}[ya(ná-**g**a)wa], which could otherwise be optimal, is eliminated due to a violation of LINKFTSTEMS. Exhaustive footing with two consecutive feet leads to unaccentedness, as in the winner (e) ^{0}[(yana)-(**g**awa)].

I have shown that the high-ranking of LINKFTSTEMS, which causes exhaustive footing, derives unaccentedness in surnames with rendaku. For thoroughness, I give two other tableaux in in (57) and (58) in order to demonstrate that the introduction of the new constraint does not affect the accentuation pattern in surnames with no rendaku application. Note that the constraint RENDAKU is ranked below IDENT(voice) (because of phonological and lexical factors associated with these surnames).

(57) Antepenultimacy in a 3μ surname with no rendaku

/saka-ℜ-ta/	IDENT(voice)	RENDAKU	NOLAPSE	LINKFTSTEMS	FTBIN	NONFIN(FT')	RIGHTMOST	WDACC	INIFT	PARSE-σ
a. ^{3}(sáka)-da	*!			*						*
b. ^{0}(saka)-(da)	*!				*			*		
c. ^{1}(saka)-(tá)		*			*!	*				
d. ^{3}(sáka)-(ta)		*			*!		*			
e. ^{0}(saka)-(ta)		*			*!			*		
f. ^{2}sa(ká-ta)		*				*!			*	*
g. ^{0}(saka)-ta		*						*!		*
☞ h. ^{3}(sáka)-ta		*								*

(58) Antepenultimacy in a 4μ surname with no rendaku

/yama-ℜ-kawa/	IDENT(voice)	RENDAKU	NOLAPSE	LINKFTSTEMS	FTBIN	NONFIN(FT')	RIGHTMOST	WDACC	INIFT	PARSE-σ
a. ^{3}ya(má-ga)wa	*!			*					*	**
b. ^{0}(yama)-(gawa)	*!							*		
c. ^{2}(yama)-(káwa)		*				*!				
d. ^{4}(yáma)-(kawa)		*					*!			
e. ^{0}(yama)-(kawa)		*						*!		
☞ f. ^{3}ya(má-ka)wa		*							*	**
g. ^{4}(yáma)-kawa		*	*!							**

As can be seen, LINKFTSTEMS does not interfere in the foot structure of a surname if rendaku voicing is not present. Antepenultimacy thus arises with a trochaic foot and a final extrametrical syllable, as we have already seen above.

Summarizing the analysis, once we posit that the application of rendaku requires a specific foot structure, namely, parsing of each element, due to the high-ranked constraint LINKFTSTEMS,

unaccentedness becomes the "default" for surnames with rendaku voicing. On the other hand, if rendaku is not realized, antepenultimacy still shows up as the basic accent pattern. This provides an answer to the long-standing problem of the correlation between rendaku realization and accentedness. Unaccentedness in surnames with rendaku is rooted in the fact that the voicing alternation is an operation for compounds, which demands compound-like prosody for its realization. The correlation can thus be viewed as a tension between two forces: surnames must behave like regular single stems with a trochaic foot and extrametricality, which derives antepenultimate accent, but rendaku realization must occur with a compound-like foot structure with two consecutive feet, which leads to unaccentedness.

For clarification, I state that the realization of rendaku requires a "compound-like" foot structure, but compound surnames are still not considered to be real compounds in the grammar. If they were actual compounds, they could further undergo compound accentuation, which places accent near the boundary (see Section 2.4.1 above). It is actually crucial that compound surnames with rendaku still undergo the accentuation process of stems, which derives unaccentedness from two consecutive feet, and not that of regular compounds.[49] The analysis is thus compatible with the original proposal: compound surnames are treated as stems and stem-phonology applies. Furthermore, this answers the question of why the correlation between rendaku and accent is found only in surnames and not in regular compounds (see Alderete 2015; also see Section 2.4.2). Regardless of rendaku application, regular compounds must parse each element into a foot (due to high-ranked LEXFT) and further undergo different requirements for accent placement. The correlation arises only when there is a tension between the regular antepenultimacy accentuation with a trochaic foot and extrametricality on one hand and unaccentedness with exhaustive footing on the other, as in the case of surnames.

[49]In this sense, compound surnames with rendaku are quite similar to truncated compounds and etymological compounds discussed in Ito and Mester (2016) (see above). They also have a compound-like foot structure, but they are still dealt with by the stem grammar and are not subject to the usual compound accentuation.

3.4.5 Deriving the five common accent patterns

I have shown that the proposed analysis readily captures the rendaku-accent correlation; the grammar correctly derives antepenultimate accent when there is no rendaku and unaccentedness when rendaku voicing is present both in three and four mora surnames. What about the other patterns that are attested? The accentuation data in surnames shown in (52) and (53) above are presented again in (59) and (60). In order to better understand the overall tendencies, I merged the surnames showing variable rendaku/accent patterns with those showing no variation by distributing the numbers of the former equally to the cells of the latter.[50] Each rendaku-accent type is supplemented with an example of a surname with that pattern.[51] The numbers that are relatively high in each surname group are shown in boldface.

(59) Frequency distribution of /ta/-surnames (based on Sugito 1965)

	Accented			Unaccented		
No rendaku	(w)	**108.5**	^{3}sáka-ta	(x)	23.5	^{0}naka-ta
Rendaku	(y)	**101.5**	^{3}háma-**d**a	(z)	**128.5**	^{0}yama-**d**a

(60) Frequency distribution of /kawa/-surnames (based on Zamma 2005)

	Accented			Unaccented		
No rendaku	(s)	**19.75**	^{3}yamá-kawa	(t)	3.75	^{0}isi-kawa
Rendaku	(u)	1.75	^{3}naká-**g**awa	(v)	**11.75**	^{0}yana-**g**awa

What is remarkable is the number in cell (y) 101.5, which shows that rendaku surnames with E2-/ta/ can be accented, going against the general correlation pattern. Their number is even comparable to those following the correlation, such as (w) 108.5 and (z) 128.5, which I have claimed to be the main patterns. What is also interesting is that an abundance of the Rendaku-Accented type

[50]For example, for /ta/-surnames, there are 56 surnames showing rendaku voicing with variable accent (accented or unaccented). I added 28 (56 divided by 2) to both the Rendaku-Accented cell and the Rendaku unaccented cell. There are 22 surnames showing variable rendaku and variable accent. I added 5.5 (22 divided by 4) to all the four cells.

[51]The name ^{0}[naka-ta] can also be realized as ^{0}[naka-**d**a] with rendaku, but is used as an example of the "No rendaku-Unaccented" class since the non-rendaku form is more common than the rendaku form.

is observed only in /ta/-surnames and not in /kawa/-surnames; see the relatively small number of (u) 1.75. Does this simply mean that surnames with /ta/ show a weaker correlation than those with /kawa/? If so, should the difference be explained merely in terms of the lexical property of each second element, as is suggested by previous studies (see Zamma 2005)? Instead of relying on such lexical specifications,[52] I claim that there is an underlying structural reason for the difference. More precisely, it is the phonological length (trimoraicity) of /ta/-surnames which causes the variability in accentedness when rendaku voicing is present.

In order to capture the variation, I assume that some of the constraints can be variably ranked, or there are subgrammars with slightly different constraint rankings à la Ito and Mester (2016). I have claimed earlier that the application of rendaku requires a compound-like foot structure (in that each element must be footed) and proposed the constraint LINKFTSTEMS as a formalization of this requirement. This constraint, however, should in theory be violable, just like any other OT constraint. If there are constraints that compete with LINKFTSTEMS, those that are ranked higher will be more important. I claim that a candidate with a violation of LINKFTSTEMS, as in ^{3}[(háma)-**d**a] 'shore-paddy,' actually surfaces as the optimal output in some circumstances. As is shown in the tableau in (61), if we flip the ranking between LINKFTSTEMS and FTBIN (i.e. FTBIN ≫ LINKFTSTEMS) while keeping all the other rankings intact, the grammar now produces accented trimoraic surnames with rendaku voicing.

[52]This is not to say that rendaku and accent in Japanese surnames do not involve any lexical specifications. In fact, certain E2 morphemes have specific accent behaviors (as if they are specified to be accented or unaccented), as is reported by Zamma (2005) and as will be discussed later in detail. Also, the variation in the data shown above is mostly lexical. It is virtually impossible to predict which trimoraic surnames with rendaku are accented (e.g. ^{3}[háma-**d**a]) and which ones are unaccented (e.g. ^{0}[yama-**d**a]), for example. These facts suggest that a full account of the phenomena of rendaku application and accentuation need lexical specifications as part of the analysis. For more details, see discussion on lexically-specific constraints in Section 5.3.3.

(61) Antepenultimacy in a 3μ surname with rendaku

/hama-ℜ-ta/	RENDAKU	IDENT(voice)	NOLAPSE	FTBIN	LINKFTSTEMS	NONFIN(FT')	RIGHTMOST	WDACC	INIFT	PARSE-σ
a. ^{3}(háma)-ta	*!									*
b. ^{0}(hama)-(ta)	*!			*				*		
c. ^{1}(hama)-(dá)		*		*!		*				
d. ^{3}(háma)-(da)		*		*!			*			
e. ^{0}(hama)-(da)		*		*!				*		
f. ^{2}ha(má-da)		*			*	*!			*	*
g. ^{0}(hama)-da		*			*			*!		*
☞ h. ^{3}(háma)-da		*			*					*
i. ^{0}hama-da		*	*!*		*			*		***

Although rendaku should be realized between two footed stems according to LINKFTSTEMS, exhaustive parsing of a three mora name necessarily results in a unary foot, violating FTBIN which is now ranked higher; candidates (c), (d) and (e) with such a degenerate foot are thus all eliminated. The optimal candidate is the one that has one trochaic foot with extrametricality, namely, antepenultimate accent, but also shows rendaku voicing at the cost of violating LINKFTSTEMS, as in (h) ^{3}[(háma)-**d**a]. One may wonder whether it is possible to make the name unaccented by assigning no feet at all, as in (i) ^{0}[hama-**d**a]; such a candidate is eliminated by NOLAPSE violations (and also loses against the winner by violating WDACC in any case).

The story is different for four mora surnames, however. As can be seen in tableau (62), the same grammar still derives unaccentedness in the case of a four mora surname.

(62) Still unaccentedness in a 4μsurname with rendaku: FTBIN ≫ LINKFTSTEMS

/yana-ℜ-kawa/	RENDAKU	IDENT(voice)	NOLAPSE	FTBIN	LINKFTSTEMS	NONFIN(FT')	RIGHTMOST	WDACC	INIFT	PARSE-σ
a. ^{3}ya(ná-ka)wa	*!								*	**
b. ^{0}(yana)-(kawa)	*!							*		
c. ^{2}(yana)-(gáwa)		*				*!				
d. ^{4}(yána)-(gawa)		*					*!			
☞ e. ^{0}(yana)-(gawa)		*						*		
f. ^{3}ya(ná-ga)wa		*			*!				*	**
g. ^{4}(yána)-gawa		*	*!		*					**

Exhaustive footing of a four mora surname enforced by LINKFTSTEMS does not produce any unary foot, unlike in a trimoraic surname. Therefore, even if FTBIN outranks LINKFTSTEMS, the optimal prosody is still unaccentedness with two consecutive feet, as in (e) ^{0}[(yana)-(**g**awa)]. That is to say, the reranking of the two constraints does not change the result in any way in four mora surnames with rendaku voicing; if rendaku applies, they are unaccented. (Compare this with the tableau in (56) above.)

The analysis captures the difference in the variation in accentedness between rendaku surnames with E2-/ta/ and rendaku surnames with E2-/kawa/. The difference in their prosodic length leads to different foot structures, causing antepenultimacy in three mora names and unaccentedness in four mora names.[53] It should also be noted that the reranking of LINKFTSTEMS and FTBIN does not affect surnames without rendaku application, since LINKFTSTEMS is not applicable to those names. If rendaku does not apply, names are accented regardless of their prosodic length. This means that the two subgrammars with minimal reranking of LINKFTSTEMS and FTBIN produce the five main accent patterns in compound surnames: see (w), (y), (z), (s) and (v) shown in boldface in (59) and (60).

[53]Notice that this comes back to the original observation made by Ito and Mester (2016): three mora items tend to be accented while four mora items are often unaccented.

To summarize, Ito and Mester's (2016) stem accent grammar with the addition of one constraint LINKFTSTEMS and reranking of two constraints can not only account for the rendaku-accent correlation but also derive the five most common accent patterns in compound surnames.

3.4.6 Deriving rarer patterns: E2-specific constraints

As can be seen in the tables in (59) and (60), there are also rarer accent patterns. For example, three mora and four mora surnames can be unaccented despite having no rendaku voicing, as in ^{0}[naka-ta] 'center-paddy' and ^{0}[isi-kawa] 'stone-river,' and four mora surnames can be accented despite showing rendaku voicing, as in ^{3}[naká-**g**awa] 'center-river.' Deriving these rarer patterns involves the addition of a few constraints and reranking.

Let us first tackle the unaccented patterns in surnames without rendaku. Following the analysis of unaccentedness given so far, we expect the foot structures of those unaccented surnames to be ^{0}[(naka)-(ta)] and ^{0}[(isi)-(kawa)] with exhaustive footing. In the case of a four mora surname, simple reranking of two constraints can derive such prosody. Recall that Ito and Mester's (2016) unaccented system with the ranking INIFT $\gg$ WDACC predicts four mora stems to be unaccented. Although the default ranking for compound surnames is WDACC $\gg$ INIFT, I argue that the ranking is reversed in a subgrammar which generates rarer surnames.[54] This grammar is illustrated with the tableau in (63), which derives unaccented ^{0}[(isi)-(kawa)] 'stone-river.'

[54]This may also be accounted for by lexically-indexed constraints (see e.g. Pater 2007). Also see below for E2-specific foot requirement constraints.

(63) Unaccentedness in a 4μ surname with no rendaku: INIFT ≫ WDACC

/isi-ℜ-kawa/	IDENT(voice)	RENDAKU	NOLAPSE	LINKFTSTEMS	FTBIN	NONFIN(FT')	RIGHTMOST	INIFT	WDACC	PARSE-σ
a. [3]i(sí-ga)wa	*!			*				*		**
b. [0](isi)-(gawa)	*!								*	
c. [2](isi)-(káwa)		*				*!				
d. [4](ísi)-(kawa)		*					*!			
☞ e. [0](isi)-(kawa)		*							*	
f. [3]i(sí-ka)wa		*						*!		**
g. [4](ísi)-kawa		*	*!							**

However, the reranking analysis cannot simply be extended to three mora surnames. As is shown by Ito and Mester (2016), their stem accent grammar with the INIFT ≫ WDACC ranking derives antepenultimacy in three mora items. As can be seen in (64), it wrongly predicts antepenultimate accent in those rarer surnames under discussion, which we want to be unaccented.

(64) Antepenultimacy for the 3μ No rendaku-Unaccented type (wrong prediction)

/naka-ℜ-ta/	IDENT(voice)	RENDAKU	NOLAPSE	LINKFTSTEMS	FTBIN	NONFIN(FT')	RIGHTMOST	INIFT	WDACC	PARSE-σ
a. [3](náka)-da	*!			*						*
b. [0](naka)-(da)	*!				*				*	
c. [1](naka)-(tá)		*			*!	*				
d. [3](náka)-(ta)		*			*!		*			
☹ e. [0](naka)-(ta)		*			*!				*	
f. [2]na(ká-ta)		*				*!		*		*
g. [0](naka)-ta		*							*!	*
☞ h. [3](náka)-ta		*								*

To solve the problem, I propose that there are E2-specific foot requirement constraints. That

is, a constraint that requires a morpheme to be parsed into a foot is indexed with each morpheme appearing as E2.[55] These E2-parsing constraints, when operative, cause exhaustive footing of a compound surname. In the case of unaccented non-rendaku surnames with E2-/ta/, the following constraint is at work.

(65) PARSE-/ta/:

The E2-morpheme /ta/ must project its own foot.

Although the constraint may seem somewhat ad hoc by itself, there is a motivation for proposing a family of E2-specific foot requirement constraints of this type. One of the recurring problems in accentuation (as well as rendaku application) in compound surnames is that the phenomenon is very much affected by the idiosyncrasy of E2 morphemes.[56] For example, certain E2 morphemes such as /kuti/ 'entrance' almost always make the surname they occur in accented (regardless of rendaku) (e.g. ^{3}[tá-**g**uti] 'paddy-entrance,' ^{3}[yamá-**g**uti] 'mountain-entrance'), while others such as /sawa/ 'mountain stream, swamp' make the surname they occur in unaccented (regardless of rendaku) (e.g. ^{0}[oo-sawa] 'big-stream,' ^{0}[kuro-sawa] 'black-stream'). It is easy to state that the latter type of morpheme is specified to be unaccented (or more precisely, cause deaccenting) as in previous studies (see Zamma 2005). However, that seems to undermine Ito and Mester's (2016) original claim that unaccentedness is derived from a specific prosodic structure, which the analysis shown here relies on. Instead, I propose that there are constraints which require specifically those morphemes to be parsed into a foot and that such constraints are ranked relatively high in the grammar. As a result, surnames with those morphemes have exhaustive footing, which leads to unaccentedness (e.g. ^{0}[(oo)-(sawa)], ^{0}[(kuro)-(sawa)]).

Given the necessity of parsing requirement constraints for these deaccenting E2 morphemes, it seems reasonable to posit the same type of constraint for /ta/. I argue that, in one of the subgram-

[55] In Section 5.3.3, I propose that there are constraints specific to each E2 morpheme and even constraints specific to each compound. For the original proposal of such word-specific constraints, see Moore-Cantwell and Pater (2016).

[56] The same is true for accent and rendaku in regular compounds. For accent, seeKubozono (1995, 1997), and for rendaku, see Rosen (2001, 2016) among others.

mars, PARSE-/ta/ is ranked above FTBIN. The grammar now produces the correct unaccented ^{0}[(naka)-(ta)] 'center-paddy,' as is illustrated in (66) below. (In this case, the ranking of WDACC and INIFT does not matter, and I use the default WDACC ≫ INIFT ranking.)

(66) Unaccentedness in a 3μ surname with no rendaku (rare pattern)

/naka-ℜ-ta/	IDENT(voice)	RENDAKU	NOLAPSE	LINKFTSTEMS	PARSE-/ta/	FTBIN	NONFIN(FT')	RIGHTMOST	WDACC	INIFT	PARSE-σ
a. ^{3}(náka)-da	*!			*	*						*
b. ^{0}(naka)-(da)	*!					*			*		
c. ^{1}(naka)-(tá)		*				*	*!				
d. ^{3}(náka)-(ta)		*				*		*!			
☞ e. ^{0}(naka)-(ta)		*				*			*		
f. ^{2}na(ká-ta)		*			*!		*			*	*
g. ^{0}(naka)-ta		*			*!				*		*
h. ^{3}(náka)-ta		*			*!						*
i. ^{0}naka-(ta)		*	*!*			*				*	**

The high-ranking PARSE-/ta/ ensures that the morpheme /ta/ is parsed into a foot. This causes exhaustive footing of the entire name and derives unaccentedness, as in (e) ^{0}[(naka)-(ta)].

It should be noted that PARSE-/ta/ is usually ranked lower than FTBIN (i.e. in other non-rendaku subgrammars), and does not affect the regular antepenultimacy accentuation in three mora surnames with no rendaku; see (57). Also note that since the constraint is specially about parsing of /ta/, it does not affect the result in /kawa/-surnames. Incidentally, now that we have E2-specific parsing constraints, the unaccented pattern in four mora surnames without rendaku shown in (63) can also be derived by ranking the constraint PARSE-/kawa/ relatively high (crucially above WDACC) in the grammar, instead of reranking WDACC and INIFT. For our purposes here, the two analyses seem equivalent. However, in order to minimize the ranking possibilities of the general constraints on accent (see the discussion below), I argue that unaccentedness in non-rendaku surnames with /kawa/ is actually derived from the ranking PARSE-/kawa/ ≫ WDACC.

The only pattern that is left now is the 4μ Rendaku-Accented type shown in (u) in table (60); some four mora surnames are accented even though they show rendaku, as in [3][naká-**g**awa] 'center-river.' Deriving antepenultimate accent in such names poses a new challenge. Recall that the basic grammar proposed above produces unaccentedness in four mora rendaku names; see (56) and (62). The grammar thus makes the wrong prediction for surnames of the 4μ Rendaku-Accented type, as shown in (67).

(67) Unaccentedness for the 4μ Rendaku-Accented type (wrong prediction)

/naka-ℜ-kawa/	RENDAKU	IDENT(voice)	NOLAPSE	LINKFTSTEMS	FTBIN	NONFIN(FT')	RIGHTMOST	WDACC	INIFT	PARSE-σ
a. [3]na(ká-ka)wa	*!								*	**
b. [0](naka)-(kawa)	*!							*		
c. [2](naka)-(gáwa)		*				*!				
d. [4](náka)-(gawa)		*					*!			
☞ e. [0](naka)-(gawa)		*						*		
☹ f. [3]na(ká-ga)wa		*		*!					*	**
g. [4](náka)-gawa		*	*!	*						**

I have proposed that LINKFTSTEMS can be ranked lower than FTBIN as one of the ranking possibilities.[57] The violation profiles of the candidates in (67) above suggest that, if LINKFTSTEMS could further be demoted all the way below WDACC, the desired candidate (f) [na(ká-**g**a)wa] would win. This, however, does not seem to be the right move. In accordance with basic assumptions of variable constraint rankings in a single language, allowing LINKFTSTEMS to be below WDACC means that the constraints in between (i.e. FTBIN, NONFIN(FT') and RIGHTMOST) could also be variably ranked with respect to each other. (In terms of constraint weights, they have similar weights.) This produces a lot of ranking possibilities, some of which even generate

[57]This explains the abundance of the 3μ Rendaku-Accented type (e.g. [0][(háma)-**d**a]) without affecting the general unaccented pattern of the four mora rendaku surnames. See Section 3.4.5 and tableaux (61) and (62).

unattested accent patterns in surnames, such as preantepenultimate *[4][(náka)-(yama)], penultimate *[2][(naka)-(yáma)] and final *[1][(naka)-(nó)]. We thus want to keep ranking possibilities minimal.

As a possible solution to the problem, I propose to split LINKFTSTEMS by E2-morphemes. That is, there are E2-specific versions of LINKFTSTEMS, as shown in (68) and (69).

(68) LINKFOOTEDSTEMS-/ta/:
A phonological exponent of a linking morpheme must occur between E1-stem projecting its own foot and E2-/ta/ projecting its own foot.

(69) LINKFOOTEDSTEMS-/kawa/:
A phonological exponent of a linking morpheme must occur between E1-stem projecting its own foot and E2-/kawa/ projecting its own foot.

I argue that the version of LINKFTSTEMS we have seen so far is actually LINKFTSTEMS-/ta/. As shown above, it is variably ranked with FTBIN, but crucially, it is ranked above WDACC. LINKFOOTEDSTEMS-/kawa/, on the other hand, is ranked somewhat lower. By default, it is still ranked right above WDACC, but for certain rare names, the ranking can be reversed (i.e. WDACC ≫ LINKFOOTEDSTEMS-/kawa/)). This grammar now derives antepenultimate accent in a four mora surname with rendaku, as in [3][na(ká-**g**a)wa], as is illustrated in (70).

(70) Antepenultimacy in a 4μ surname with rendaku (rare pattern)

/naka-ℜ-kawa/	RENDAKU	IDENT(voice)	NOLAPSE	LKFTSTEMS-/ta/	FTBIN	NONFIN(FT')	RIGHTMOST	WDACC	LKFTSTEMS-/kawa/	INIFT	PARSE-σ
a. ^{3}na(ká-ka)wa	*!									*	**
b. ^{0}(naka)-(kawa)	*!							*			
c. ^{2}(naka)-(gáwa)		*				*!					
d. ^{4}(náka)-(gawa)		*					*!				
e. ^{0}(naka)-(gawa)		*						*!			
☞ f. ^{3}na(ká-ga)wa		*							*	*	**
g. ^{4}(náka)-gawa		*	*!						*		**

LINKFTSTEMS-/ta/ is not relevant to the foot structures of /kawa/-surnames. Since WDACC outranks LINKFTSTEMS-/kawa/, it is better to have accent at the cost of violatingLINKFTSTEMS-/kawa/, and (f) ^{3}[na(ká-**g**a)wa] is the winner.

As stated above, LINKFTSTEMS-/kawa/ usually outranks WDACC. With this ranking, the grammar still generates the default unaccented pattern for four mora surnames with rendaku. For thoroughness, this is shown in the tableau in (71) with the name of the Rendaku-Unaccented type ^{0}[(yana)-(**g**awa)] 'willow river.'

(71) Default unaccentedness in a 4μ surname with rendaku (basic pattern)

/yana-ℜ-kawa/	RENDAKU	IDENT(voice)	NOLAPSE	LKFTSTEMS-/ta/	FTBIN	NONFIN(FT')	RIGHTMOST	LKFTSTEMS-/kawa/	WDACC	INIFT	PARSE-σ
a. [3]ya(ná-ka)wa	*!									*	**
b. [0](yana)-(kawa)	*!								*		
c. [2](yana)-(gáwa)		*				*!					
d. [4](yána)-(gawa)		*					*!				
☞ e. [0](yana)-(gawa)		*							*		
f. [3]ya(ná-ga)wa		*						*!		*	**
g. [4](yána)-gawa		*	*!					*			**

This only involves minimal reranking of LINKFTSTEMS-/kawa/ and WDACC, and most importantly, WDACC is always outranked by NONFIN(FT') and RIGHTMOST, which ensures that the location of the accent is antepenult, and not any other position. The proposal is still compatible with all the other analyses of the common patterns proposed above, and does not overgenerate unattested accent patterns. One may think that the analysis in the end resorts to lexical specifications (or lexically-specific constraints) in order to capture differences between /ta/-surnames and /kawa/-surnames. Note, however, that it still does not completely undermine the claim I made earlier about their difference in general accentuation. I have shown that reranking of LINKFTSTEMS and FTBIN accounts for the variation in accentedness in /ta/-surnames with rendaku voicing (e.g. [0][yama-**d**a] vs. [3][háma-**d**a]). (See Section 3.4.5.) Although the constraint itself is now specific to the morpheme /ta/, the variation arises specifically because /ta/-surnames are three mora long (and it would not occur with four mora /kawa/-surnames).

Splitting LINKFTSTEMS according to E2 morphemes is stipulative. However, once we try to account for the accent patterns of surnames in general, it looks reasonable to say that different E2 morphemes violate the footing requirement of rendaku application expressed here by

LINKFTSTEMS in a different manner (supposing it is a legitimate constraint). As stated above, some E2 morphemes such as /kuti/ 'entrance' almost always make the surname they occur in accented while also triggering rendaku (e.g. ^{3}[yamá-**g**uti] 'mountain-entrance'). Zamma (2005) proposes that these E2 morphemes are specified to be [+accent]. This putative [+accent] feature does not seem to be linked to any of the actual prosodic structure of the input (e.g. not /kúti/, /kutí/, etc.), since it is always realized on a syllable in E1 (e.g. ^{3}[yamá-**g**uti]). A reasonable interpretation, then, would be that it is a kind of floating feature associated with E2-morphemes, and that the position it falls on is determined by the grammar. Using the stem accent grammar with general LINKFTSTEMS I proposed above, let us see whether this accent specification can correctly derive surnames that are always accented. The grammar is shown in (72). I assume that there is a high-ranked constraint like REALIZEACCENT, which requires a floating [+accent] feature to be realized in the output. Note that the LINKFTSTEMS constraint is the general version that operates on any E2 morphemes.

(72) Wrong accent patterns by a mere accent-specification

/yama-ℜ-kuti[+acc.]/	REALIZEACCENT	RENDAKU	IDENT(voice)	NOLAPSE	LINKFTSTEMS	FTBIN	NONFIN(FT')	RIGHTMOST	WDACC	INIFT	PARSE-σ
a. ^{3}ya(má-ku)ti		*!								*	**
b. ^{0}(yama)-(kuti)	*!	*							*		
☞ c. ^{2}(yama)-(gúti)			*				*				
☞ d. ^{4}(yáma)-(guti)			*					*			
e. ^{0}(yama)-(guti)	*!		*						*		
☹ f. ^{3}ya(má-gu)ti			*		*!					*	**
g. ^{4}(yáma)-guti			*	*!	*						**

As can be seen, an analysis with accent specifications in E2 runs into a problem. If there is an idiosyncratic demand for accent realization, the grammar assigns accent to a surname but still respects LINKFTSTEMS if rendaku voicing is present. As a result, four-mora surnames re-

ceive penultimate or preantepenultimate accent, a pattern which is never attested.[58] However, it seems unfavorable to completely abandon LINKFTSTEMS; it is well motivated by the raison d'être of rendaku voicing and an analysis based on the constraint captures the five main accent patterns of surnames, as we have seen above. I thus propose that those accenting E2 morphemes have no accent specification whatsoever in their underlying forms. Rather, it is low ranking of LINKFTSTEMS associated with those morphemes which cause the surnames they occur in to be always accented. For example, LINKFTSTEMS-/kuti/ is (always) ranked below WDACC, and thus rendaku surnames with /kuti/ are always accented with one trochaic foot and extrametricality, violating the prosodic requirement for rendaku application, as in [3][ya(má-**g**u)ti]. This is illustrated in (73) below. (Due to limitations of space, I omit other E2-specific LINKFTSTEMS constraints such as LINKFTSTEMS-/ta/ and LINKFTSTEMS-/kawa/, but their presence does not affect the result of surnames with /kuti/.)

(73) Capturing accentuation by the low ranking of an E2-specific LINKFTSTEMS

/yama-ℜ-kuti/	RENDAKU	IDENT(voice)	NOLAPSE	FTBIN	NONFIN(FT')	RIGHTMOST	WDACC	LKFTSTEMS-/kuti/	INIFT	PARSE-σ
a. [3]ya(má-ku)ti	*!								*	**
b. [0](yama)-(kuti)	*!						*			
c. [2](yama)-(gúti)		*			*!					
d. [4](yáma)-(guti)		*				*!				
e. [0](yama)-(guti)		*					*!			
☞ f. [3]ya(má-gu)ti		*						*	*	**
g. [4](yáma)-guti		*	*!					*		**

[58] One possible solution could be to somehow specify that the [+accent] feature in the E2 input falls on some specific position in a compound, such as on the last syllable of E1. Putting aside the question of whether such a specification is possible in phonological theory, it would miss the generalization that accent falls on the syllable containing the antepenultimate mora of the entire name. In the case of one mora accenting E2-morphemes such as /se/ 'riffle,' accent falls not on the last syllable of E1 but on the second-to-last syllable of E1 (e.g. [yáma-se] 'mountain-riffle').

As we see in the tableau, if LINKFTSTEMS for the E2 morpheme of the surname is ranked low, the grammar assigns the default antepenultimate accent. This analysis based on E2-specific LINKFTSTEMS accounts for the non-variability of the location of accent in surnames. The rankings of the basic constraints on accent, specifically NONFIN(FT'), RIGHTMOST and WDACC, are fixed, and the grammar only generates antepenultimate accent. Surnames with certain E2 morphemes are always accented even with the presence of rendaku voicing, not because they are specified to receive accent by lexical specifications in E2, but because they are not subject to the parsing requirement of rendaku and therefore undergo the default accentuation process.

The analysis can actually be extended to the accent patterns of non-rendaku surnames. There are E2 morphemes that almost never undergo rendaku and (almost) always cause the surname they occur in to be accented. Such morphemes include /sita/ 'bottom,' /tani/ 'valley,' /se/ 'riffle,' /saka/ 'slope' and so on (Zamma 2005).[59] Their accent behavior is already expected from the default constraint rankings which capture the rendaku-accent correlation; if no rendaku, accented. What remains to be explained is the fact that unaccentedness (almost) never arises in those surnames, unlike in surnames with /ta/ and /kawa/ which may occasionally be unaccented without rendaku (e.g. ^{0}[naka-ta], ^{0}[isi-kawa]). I argue that the reason lies in the rankings of lexically-specific constraints. Recall that I have proposed E2-specific parsing constraints such as PARSE-/ta/ and PARSE-/kawa/ which could cause unaccentedness in non-rendaku surnames by requiring E2 morphemes to form an independent foot (e.g. ^{0}[(naka)-(ta)], ^{0}[(isi)-(kawa)]). I argue that PARSE constraints specific to the accenting E2-morphemes given above are ranked especially low and they never cause surnames with those morphemes to be unaccented. Again, accented surnames are not specified to be accented; they simply show the default antepenultimate pattern.

The analysis with E2-specific constraints proposed here suggests that a language learner must face a lot of constraints. As will be discussed in Chapter 5, such constraints are actually necessary in order to learn and produce the rendaku and accent patterns of surnames. As will be

[59] e.g. ^{3}[yamá-sita] 'mountain-bottom,' ^{3}[naká-tani] 'center-valley,' ^{3}[íwa-se] 'rock(y)-riffle,' ^{3}[aká-saka] 'red-slope'; a few exceptions with variable pronunciations are ^{0}[mizu-tani] ~ ^{3}[mizú-tani] 'water-valley,' ^{0}[oo-saka]~^{3}[óo-saka] 'big-slope.'

shown, however, having a lot of lexically-specific constraints in a grammar is not equal to lexicalizing/memorizing all the patterns. Learners can still learn phonological patterns from data with such lexically-specific constraints and produce the patterns productively.

In summary, I have shown that, with the addition of lexically-specific constraints and reranking of some constraints, the stem accent grammar can generate rarer patterns in the data. I have claimed that E2 morphemes have no specifications of accent and their accentuation behaviors are derived from interactions of constraints, including lexically-specific ones.

3.4.7 Summary

In this section, I have pursued the proposal that compound surnames are subject to the phonology of stems. I have given an account of their prosodic patterns and how they relate to rendaku application. I have argued that compound surnames behave like regular single stems in terms of accentuation, which explains their predictable antepenultimate accent patterns. I have also addressed the issue of the correlation between accentedness and rendaku application. I have argued that the correlation arises from the tension between a pressure for having a stem-like foot structure, which would derive the default antepenultimacy, and a pressure to have a compound-like foot structure when compound voicing is present, which would derive unaccentedness. I have shown that Ito and Mester's (2016) account of accent in Japanese stems can be extended to accent of compound surnames and further derives the rendaku-accent correlation.

3.5 Remarks on other patterns

3.5.1 Remarks on E1 and E2 length effects

In Chapter 2, I have presented another prosodic factor affecting rendaku in surnames, namely, the length of the first element. Generally speaking, rendaku is more likely to apply when E1 is monomoraic or trimoraic than when it is bimoraic. Can this prosodic factor also be derived from the fact that compound surnames are treated as single stems in the grammar? Analyzing

the E1-length effects under stem-phonology would actually be problematic. As mentioned above, the same prosodic-size effects are found in regular compounds (see Section 2.4.3; also see Rosen 2001, 2003; Irwin 2014a, 2016a). This indicates that the patterns are driven by some grammatical principle which is more general than stem-internal phonology. Developing a full account of the E1-length effects would require more detailed analyses of the prosody and rendaku in both regular compounds and compound surnames, which is beyond the scope of this dissertation. In what follows, I will simply accept as true the description of the E1-length effects given in the previous studies,[60] but leave a formal analysis of the patterns for future research.[61]

3.5.2 Remarks on lexical propensities

One other factor of rendaku which has been raised in the previous chapter is the lexical propensities of E2-morphemes and those of compounds themselves (see Section 2.5). It has been observed that certain morphemes are generally more likely (or less likely) to undergo compound voicing than others when they appear as E2 of a compound surname (Zamma 2005). In addition, although phonological factors and E2's lexical propensities each contribute to the likelihood of voicing, whether a given surname actually undergoes rendaku or not is also affected by the idiosyncratic properties of that compound. Such lexically-conditioned patterns are not unique to rendaku in surnames; regular compounds are also known for showing the same kind of lexical irregularities in rendaku application (see e.g. Martin 1987; Vance 1979, 1980b; Rosen 2001, 2016; Irwin 2014a, 2016a).

This study does not intend to offer an explanation as to why individual E2-morphemes or individual surnames have the particular lexical properties with respect to rendaku that are associated to them. Take, for example, the surnames /taka-ta/ [taka-da] 'high-paddy' and /saka-ta/ [saka-ta] 'slope-paddy'; even though the two are underlyingly quite similar, the former usually shows rendaku while the latter does not. The difference seems to be based on purely lexical fac-

[60] Indeed, the effects are confirmed in a corpus study of surnames (see Section 4.2).

[61] Rosen (2003) for an attempt to formally analyze the effects of trimoraic E1 in rendaku in regular compounds.

tors, and it is beyond the scope of phonological research to spell out why their voicing patterns are the way they are. Instead, this dissertation is concerned with *how* such lexically-conditioned rendaku patterns should be treated in phonological theory. In Chapter 5, I will give more discussion on what problems are posed by lexically-conditioned rendaku application and propose a grammar model with lexically-specific constraints (Moore-Cantwell and Pater 2016) of the kind shown in the analysis of the accent patterns. I will further show that the proposed model successfully captures the lexical irregularities of rendaku in surnames and claim that lexical factors should be incorporated into the grammar rather than into the lexicon.

3.6 Chapter summary

In this chapter, I have given an account of the phonological patterns of Japanese compound surnames. I have proposed that the grammar treats surnames as single stems rather than compounds due to their semantic non-compositionality, and that rendaku application and accentuation in surnames are governed by stem phonology. The proposal gives coherent explanations for the issues raised in Chapter 2. Under the proposed analysis, the effects of E1's last consonants such as Strong Lyman's Law reported in the literature can be restated as the application of stem-internal phonological restrictions. I have also shown that the predictability of accent location and the correlation between rendaku and accentedness in compound surnames can be derived by extending Ito and Mester's (2016) analysis of accent of Japanese stems. The proposal has made further predictions about rendaku application, which will be tested in the next chapter.

CHAPTER 4

Corpus study and experiment

4.1 Chapter overview

This chapter is concerned with a corpus study of existing surnames and a judgment experiment using non-existing surnames. Section 4.2 presents the corpus study results which reveal that rendaku application in real surnames exhibits patterns that are similar to the phonotactic patterns of single stems, supporting the claim that stem phonology applies to compound surnames. The results will be compared to the rendaku patterns in regular compounds. Section 4.3 shows that, in experimental settings, Japanese speakers apply rendaku to nonce surnames in the way similar to what is seen in the corpus data, providing evidence for the productivity of the patterns. The section also gives a comparison of the corpus results and the experimental results, and discusses their similarities and differences.

4.2 A corpus study: Rendaku in existing surnames in social media

This section presents the results of a corpus study of existing surnames. I collect the readings of surnames appearing in social media and calculate the rendaku rate of each name. The hypotheses about segmental factors in rendaku proposed in the previous chapter will be tested against the data. I will also discuss differences between the rendaku patterns in compounds surnames and those in regular compounds.

4.2.1 The aims of the study

In Chapter 3, I have claimed that the peculiar rendaku patterns in surnames can be attributed to the application of stem-internal phonology. The proposed hypothesis makes several predictions about how combinations of segments in the first element (E1) and the second element (E2) of compound surnames will promote or inhibit the application of rendaku. In order to test these predictions, a new, objective and extensive examination of the voicing alternation in surnames is necessary.

The data reported in previous studies on rendaku in surnames are, by and large, inadequate to fully address the issues raised in the preceding chapters.[1] Most of the studies are only concerned with surnames with certain E2 morphemes such as /ta/ 'paddy, rice field.' Analyzing names with a limited number of E2 morphemes does not provide an overall picture of the rendaku patterns of Japanese surnames. As stated above, the place of E2's initial obstruent (/s/, /t/, /k/ or /h/) may exert a different effect on rendaku application depending on its combination with E1's last consonant, and each E2 morpheme itself may also show an idiosyncratic behavior. The most extensive work is Zamma (2005) who reports judgment data of rendaku in surnames with seventeen different E2 morphemes obtained from five native speakers of Japanese. The study, however, still does not report the details of rendaku application based on the place of E1's last consonant and E2's initial consonant, which would make it possible to address the effect of Identity Avoidance.

Also, the observations of previous studies are mostly based on the intuitions of a limited number of speakers, sometimes including authors themselves. Linguists' judgments could be biased by their own theory. Having a small number of consultants also makes it difficult to describe variabilities in phonological patterns in detail. As is shown in (6) above, the surname /naka-ta/

[1]Besides Sugito (1965), Kubozono (2005) and Zamma (2005), linguistic studies which discuss rendaku patterns in surnames include Kindaichi (1976/2005); Nakagawa (1978); Sato (1989); Hirata (2010, 2011); Ohta (2013); Asai (2014); Vance and Asai (2016); also see Ito and Mester (2003:109) and Kawahara (2015b:477–479). Sugito's (1965) data are concerned with rendaku judgments on surnames with /ta/ as E2 obtained from six speakers. Kubozono (2005) uses Sugito's data for his analysis. Zamma (2005) uses judgments of five speakers on surnames with seventeen E2 morphemes. Hirata (2010, 2011) collects the readings of trimoraic surnames with E2-/ta/ mainly from dictionaries listing historical figures. Ohta (2013) describes rendaku and accent in surnames with one E2 /kawa/ based on his own intuitions. Asai (2014) collects judgments from twenty speakers using surnames with two E2 morphemes /saki/ and /taki/ as the stimuli. Other studies only report a small amount of surname data collected on their own or taken from the previous literature.

'center-paddy' (with a /k...t/ sequence across the boundary) can be pronounced either [naka-ta] or [naka-da], but the non-rendaku form is more common than the rendaku form. On the other hand, another variable surname /tuka-kosi/ 'mount-crossing' (with a /k...k/ sequence) surfaces more often as the rendaku form [tuka-**g**osi] than as the non-rendaku form [tuka-kosi].[2] Notice that the patterns of variation in these names are what is expected given Identity Avoidance effects; a /k...t/ sequence is already dissimilar and voicing is not mandatory, while an identical /k...k/ sequence promotes rendaku for the purpose of dissimilation. As these examples suggest, the rendaku rates of variable names, when taken as a whole, could tell us how E1's last consonant affects the likelihood of rendaku application. Additionally, none of the previous studies conducts statistical analyses of the data mainly due to the limited numbers of consultants and sample size. All of the proposed generalizations are thus still in need of validation.

To address all these issues, I create a corpus of Japanese surnames with rendaku application rates, collecting the pronunciations of names from social media. The large-scale data enable us to assess comprehensively and quantitatively the validity of the generalizations given in previous studies as well as the hypotheses proposed in this dissertation.

4.2.2 Methods

As a first step, I created an original list of Japanese surnames by combining data obtained from two existing databases. Main data came from *the Database of Japanese Surnames and the Rankings* (Shirooka and Murayama 2011) which lists 25,000 surnames in kanji (Chinese characters) and the number of households having each surname which are registered in telephone directories. From the database, all the surnames written with two characters (i.e. bimorphemic surnames) were extracted. Since kanji can have multiple different readings, possible pronunciations of each surname were taken from another online database (Suzaki 2013) that provides possible readings of surnames.

Note that the created list still does not give information as to whether or how often each surname

[2] According to the results which will be presented below, for /naka-ta/ the ratio is 80.4% non-rendaku ([naka-ta]) and 20.6% rendaku ([naka-**d**a]), and for /tuka-kosi/ it is about 81.7% rendaku ([tuka-**g**osi]) and 18.3% non-rendaku ([tuka-kosi]).

undergoes rendaku. As mentioned above, the kanji script does not reflect rendaku voicing, and the household numbers of surnames in the database are all based on kanji. For example, it gives 39,637 as the number of registered households having the surname written "中田" (underlyingly /naka-ta/ 'center-paddy'; but also see below for other possible underlying forms); however, the name can read [naka-ta] or [naka-da], and the ratio of the rendaku form to the non-rendaku form is not available. To solve the issue, I further compiled data from two social networking services: *mixi* and *Facebook*. More particularly, I collected the pronunciations of users' surnames from these websites and calculated the rendaku rate of each surname based on the counts.

Mixi (stylized as "mixi") is a popular social networking website in Japan with about 13 million users as of 2012. Many mixi users make their names public so that other users can search for them. Some users also provide their names in kanji and the readings in phonographic scripts such as hiragana, katakana and the Roman alphabet, which always indicate rendaku voicing. A friend search function allows us to search for users whose profiles contain a particular kind of information. For example, one can type someone's surname in kanji and its reading in hiragana with a space in-between in the search box, as in «中田 nakata».[3] (The hiragana script is replaced here by the Roman alphabet for expository purposes. The actual search term used was «中田 なかた»). The function then returns the number of hits, namely, the number of users who have the search terms in their user profiles. A computer script was used to search for the rendaku reading and the non-rendaku reading of each surname in the list. For example, for /naka-ta/ 中田, the script searched for «中田 nakata» and «中田 nakada». From the obtained numbers, the rendaku rate of the surname was calculated. As there were 85 users with the keyword «中田 nakata» and 22 users with the keyword «中田 nakada», the rendaku rate of this surname was deemed to be 22/(85+22) or 20.56% according to the kanji-hiragana searching method on mixi. The procedure was repeated using a combination of kanji and the other two phonographic scripts: that is, the kanji-katakana combination and the kanji-alphabet combination. The results for /naka-ta/ using these methods

[3]Search terms are put in guillemets.

were 18.00% and 18.18% respectively.[4]

Data were also collected from Facebook in a similar way. Many Japanese users of the service register their names both in kanji and in the Roman alphabet. Since Facebook's built-in friend search function does not give the exact number of users with a search term, I used a website-internal search function of the search engine *Google*. One can type a search term (e.g. "XYZ") with a specification of a target website in the search box, as in «XYZ site:facebook.com». Google then returns the number of hits, namely, the number of webpages containing the term inside facebook.com. Using this function, I collected the numbers of the rendaku form and the non-rendaku form of each surname in the list appearing in Facebook. For search terms, I used surnames written in kanji and their readings written in the Roman alphabet, as in «中田 "nakata" site:facebook.com» and «中田 "nakada" site:facebook.com».[5] Again, a computer script was used to automate search actions, and the rendaku rate of each surname was calculated from the obtained numbers. As there were 50,800 hits with the keyword «中田 nakata» and 17,500 hits with the keyword «中田 nakada», the rendaku rate of this surname was deemed to be 17,500/(50,800+17,500) or 25.62% based on the kanji-alphabet searching method on Facebook.

The final rendaku rate of each surname was then obtained by averaging the results of the different searching methods. In the case of /naka-ta/ 中田, I concluded its rendaku rate to be 18.91%, by averaging 20.56% (the kanji-hiragana method on mixi), 18.00% (the kanji-katakana method on mixi), 18.18% (the kanji-alphabet method on mixi) and 25.62% (the kanji-alphabet method on Facebook). Using the same calculation procedure, I concluded the rendaku rate of /naka-hara/ 中原 'center-field' to be 0.05% (that is, almost never rendaku), that of /naka-kawa/ 中川 'center-river' to be 99.2% (almost always rendaku) and that of /naka-sima/ 中島 'center-island' to be 66.65% (variable with the rendaku form being somewhat more common). These rendaku rates were further

[4] /h/ may alternate with [w] in certain surnames, as in /huzi-hara/ → [huzi-wara] 'wisteria-field.' For names with [h]-initial E2, the number of the form with [w] was also collected as a non-rendaku realization of the name.

[5] The readings were put in double-quotations in order to avoid auto-corrections by Google. Without quotations, if either the rendaku form or the non-rendaku form is too infrequent, the engine corrects the search and shows the result of the other (more frequent) form. For example, /yama-ta/ 山田 'mountain-paddy' always undergoes rendaku to be realized as [yama-**d**a]. If «山田 yamata site:facebook.com» without quotations is used for a search, the result rather shows the numbers of pages containing the term «山田 yamada» by correction.

used to calculate the actual household number of each of the rendaku and non-rendaku forms. For /yama-saki/ 山崎 'mountain-cape' which has the rendaku rate of 76.26% and the total households of 121,564, I concluded 92,705 of them have the rendaku reading [yama-**z**aki] and 28,859 of them have the non-rendaku reading [yama-saki].

There is one potential problem in calculating the household number of each possible form of a surname in this manner. Besides the reading variation based on rendaku application (e.g. [naka-ta]~[naka-da]), a surname written in kanji may have yet other kinds of reading possibilities. For example, the E1 morpheme of a surname may have multiple readings. In the case of the surname 中田, which I have shown as being underlyingly /naka-ta/, the kanji of the first element also has the Sino-Japanese reading /tyuu/. (/naka/ is a native reading.)[6] The surname can thus also be represented as /tyuu-ta/, which may in turn be realized as [tyuu-ta] or [tyuu-da].[7] Similarly, the E2 morpheme of a surname may have multiple readings. The morpheme 谷 'valley' has two native readings /tani/ and /ya/. For the surname 中谷 'center-valley,' we can posit two underlying forms: /naka-tani/ (potentially realized as [naka-tani] or [naka-**d**ani]) and /naka-ya/ (always realized as [naka-ya]; rendaku is not relevant since E2's initial consonant is a sonorant).[8] In this study, if a kanji-written surname has different forms of this kind (which are not dependent on variation in rendaku), I consider them to be underlyingly different names. However, the original kanji-based database of surnames (Shirooka and Murayama 2011) treats them the same and gives one single total household number. In other words, the household number of a surname is unknown if it shares kanji with other surnames which are phonologically distinct.

In order to obtain the accurate household numbers of underlying surname forms, I also calcu-

[6] As stated above, one Chinese character may have multiple readings: one or more native Japanese readings which correspond to the pronunciations of the native Japanese words with the meaning of that character (e.g. 人 /hito/ 'person') and often several Sino-Japanese readings which correspond to the pronunciations of the same word borrowed at different times of the history (e.g. 人 /nin/, /zin/ 'person').

[7] Narrower transcriptions would be [tɕɯː-ta] and [tɕɯː-da]. See Section 1.3.2 for notes on transcriptions.

[8] /tyuu-tani/ and /tyuu-ya/ are also logically possible underlying forms of the surname 中谷. However, since Suzaki (2013) does not give any of these forms as possible readings of the surname, they were excluded from the search. /tyuu-den/, another logically possible underlying form for 中田, was excluded for the same reason.

lated the UR rates of each kanji-written surname. As discussed above, according to the Facebook searches conducted for the surname /naka-ta/ 中田, the readings [naka-ta] and [naka-da] had 50,800 hits and 17,500 hits respectively. According to the searches conducted for /tyuu-ta/ 中田, [tyuu-ta] and [tyuu-da] had 165 hits and 334 hits respectively.[9] Since there are 68,300 (50,800+17,500) hits in total for /naka-ta/ and 499 (165+334) hits in total for /tyuu-ta/, their UR rates were deemed to be 99.27% (= 68,300/(68,300+499)) and 0.72% (= 499/(68,300+499)) respectively. Based on these UR rates, the household number of each surname UR was calculated; of the total 39,637 households of 中田, 39,350 are /naka-ta/ and 287 are /tyuu-ta/. For /naka-tani/ and /naka-ya/ which are both written 中谷, their UR rates turned out to be 77.40% and 22.60% respectively. I thus concluded that, of the total 19,909 households of 中谷, 15,410 are /naka-tani/ and 4,499 are /naka-ya/.

Finally, for surnames that are potential rendaku undergoers, these UR household numbers were used to calculate the number of the rendaku form and the non-rendaku form. For instance, for /naka-ta/ 中田 with the total UR households of 39,350 and the rendaku rate of 18.91%, 8,102.53 (= 39,350*18.91%) households have the rendaku reading [naka-da] and 31247.47 (= 39,350–8,102.53) have the non-rendaku reading [naka-ta]. For /naka-tani/ 中谷 with the UR households of 15,410 and the rendaku rate of 0.14%, 21.55 (= 15,410*0.14%) households have the rendaku reading [naka-dani] and 15388.45 (= 15410–21.55) have the non-rendaku reading [naka-tani].

From the completed list of surname URs, those that had the second element with initial /s/, /t/, /k/ or /h/ (i.e. potential rendaku undergoers) were extracted. Infrequent surnames with fewer than 1,500 households were excluded from analysis. Those with a voiced obstruent in E2 (e.g. /taka-sugi/ 'tall-cedar'; /matu-huzi/ 'pine-wisteria') were excluded since rendaku was expected to be blocked categorically due to the application of normal Lyman's Law. Certain surnames have genitive /no/, /na/ or /ga/ between E1 and E2, which is not reflected in the orthography

[9]In conducting searches in the Roman alphabet, the surname readings were spelled according to the Hepburn Romanization system, as in «中田 "chuta" site:facebook.com» and «中田 "chuda" site:facebook.com». Although the reading [tyuu-da] is by far rarer than [naka-ta] and [naka-da], it does exist as a regional variant of 中田 mainly in the Okinawa Islands. The hits of [tyuu-ta], on the other hand, may be due to some noise. (See below for the caveat on the searching methods.)

(e.g. /ki-no-sita/ 木下 'tree-Gen.-bottom'; /wata-na-he/ 渡辺 'crossing-Gen.-edge'); they were excluded since genitive particles and rendaku voicing usually do not cooccur (see Lyman 1894; Vance 2005b). Those with numeral E1 possibly followed by a classifier (e.g. /mi-tu-hasi/ 三橋 'three-Classifier-bridge'; /mi-sima/ 三島 'three-island') were also excluded since numerals generally inhibit rendaku application in regular compounds (see Nakagawa 1966; Irwin 2012). Those with a Sino-Japanese morpheme as E2 were excluded since rendaku is generally less likely to apply to non-native E2 morphemes (see Lyman 1894; Martin 1952:48; Otsu 1980; Ito and Mester 1986; Vance 1996; Ito and Mester 2003:144–153; Irwin 2005).[10,11] Surnames with Sino-Japanese E1, on the other hand, were included in the data since the stratum membership (whether native or not) of E1 generally does not affect the applicability of rendaku in regular compounds (see Ito and Mester 2003:144–153; Ohno 2000:155; but cf. Tamaoka et al. 2009:30–31). Some surnames could be analyzed as being dvandvas (e.g. /yama-kawa/ 'mountain-river'; the possible interpretations are 'mountain river' and 'mountains and rivers'), which usually do not undergo rendaku; however, no data were excluded for this reason since it is essentially impossible to distinguish ambiguous semantic structures of a surname (see Section 3.2.1). The resulting list contained 1,064 surnames with 122 distinct E2 morphemes.

Several caveats should be mentioned. First, the corpus study is entirely based on text and no information about the accent patterns of surnames were gathered. For this reason, the hypotheses on accentuation will not be addressed. Second, the study does not take possible regional differences

[10]Some surnames with Sino-Japanese morphemes do undergo rendaku, commonly when E1's last consonant is a moraic nasal (e.g. /an-too/ → [an**d**oo] 'peaceful-wisteria' (derived from Fujiwara through truncation-compounding; see Section 2.6.3); /an-sai/ → [an**z**ai] 'peaceful-West').

[11]Another problem with surnames (or compounds in general) with Sino-Japanese E2 morphemes is that it is sometimes impossible to determine whether or not voicing of E2 is a case of rendaku application. One Sino-Japanese morpheme may have multiple pronunciations due to the fact that the same morpheme was borrowed from Chinese successively at different times of the history. Some were borrowed as having a voiced obstruent at some point but were later reborrowed as having a voiceless obstruent, as in /ti/ and /di/ for 地 'earth, land.' There are surnames with such an E2 morpheme (e.g. [miya-ti]~[miya-di] 'shrine-land'), but it is unclear whether they are voiced as a result of rendaku application or already voiced underlyingly. (See Okumura 1952; Vance 1996; Irwin 2005 for similar cases in regular compounds.)

in the pronunciations of surnames into account.[12] Since the readings of surnames were taken from any pages in mixi and Facebook without a specification of where the surnames were from, the compiled data represent the distributions of the readings of surnames in Japan as a whole. Third, the data collected from social media may contain some noise. For example, pages in Facebook can contain not only the names of individuals but also the names of cities, regions, organizations, companies and so on. It is possible that the rendaku rates of surnames have been affected by such information of proper nouns other than surnames. Lastly, Google may show slightly different hits every time a search of the same term is conducted. Those differences are considered to be minor and not affect the overall results significantly.

Data with apparent errors were either discarded or corrected by hand.[13] There were a handful of cases where one or two of the four searching methods showed significantly different results from those of the other methods. In such cases, the numbers which looked most reliable to the author were adopted. Some searches did not get enough hits to calculate rendaku rates. The results were not used if the sum of the hits was fewer than five. Hit results obtained from searches in mixi were generally small in number. For this reason, I discarded almost all the results of infrequent surnames with fewer than about 4,000 households taken from mixi.

[12]There are informal reports that, as far as certain surnames with rendaku variation such as /yama-saki/ 'mountain-cape' and /naka-sima/ 'center-island' are concerned, the rendaku form is more commonly found in the Western regions of Japan than in the Eastern regions. (See e.g. Iwasaki 2013:42; Morioka 2011:221.) Since it is not viable to validate this with the data at hand, I will leave the issue open.

[13]The most common errors were based on the spelling of long vowels. A version of the Hepburn romanization system commonly used for proper nouns, which was employed for searches in the study, does not usually represents vowel length contrasts. For example, the common E1 morephemes /o/ 'small' and /oo/ 'big' are both spelled with a single "o." (Some people use the spelling "oh" to represent /oo/, but the use is not consistent.) The surname /o-sawa/ 'small-stream' always undergoes rendaku, as in [o-**z**awa], and /oo-sawa/ 'big-stream' never does, as in [oo-sawa]; but the differences could not be detected in the corpus study since the rendaku form and the non-rendaku form of both surnames were spelled «osawa» and «ozawa». Note, however, that these surnames do not have a consonant in E1's final syllable and thus the errors do not affect the main predictions proposed in the dissertation.

4.2.3 Predictions

Let us review the main predictions made by the hypothesis that rendaku in surnames is governed by stem-internal phonotactics. I have proposed that stem-bounded Lyman's Law (Lyman 1894) applies to the whole word in the case of a compound surname (unlike in a regular compound), deriving so-called Strong Lyman's Law effects. In other words, a voiced obstruent in E1 inhibits application of rendaku on E2's initial segment across the boundary between E1 and E2, as shown in (74).

(74) E1-voiced obstruents block rendaku

a.	naga	+	sima	→	naga-sima	*naga-**z**ima	'long-island'
b.	siba	+	ta	→	siba-ta	*siba-**d**a	'grass-paddy'
c.	huzi	+	kawa	→	huzi-kawa	*huzi-**g**awa	'wisteria-river'
d.	sugi	+	hara	→	sugi-hara	*sugi-**b**ara	'cedar-field'

Blocking of rendaku by a voiced obstruent in E1 has been reported to occur in surnames with /ta/ as E2 (Sugito 1965) and other common surnames (Zamma 2005). We expect such patterns to hold systematically in a large corpus. In what follows, I use the term "Strong Lyman's Law" to refer to the phenomenon of boundary-spanning Lyman's Law application. It should be noted, however, that the law itself is nothing but the normal stem-bounded Lyman's Law. (See Section 3.3.1 for more details.)

The hypothesis also predicts that stem-internal restrictions on sequences of homorganic voiceless obstruents (Kawahara et al. 2006) apply to the whole compound surname and affect rendaku application, deriving so-called Identity Avoidance effects. For example, the voicing alternation will be triggered when a voiceless obstruent in E1's last mora and a voiceless obstruent in E2's initial mora share place, as shown in (75).

(75) Similar obstruent sequences trigger rendaku

a.	nisi	+	sima	→	nisi-**z**ima	*nisi-sima	'West-island'
b.	matu	+	ta	→	matu-**d**a	*matu-ta	'pine-paddy'
c.	naka	+	kawa	→	naka-**g**awa	*naka-kawa	'center-river'

Descriptions given by previous studies (Kubozono 2005; Zamma 2005) suggest that surnames with E2-/ta/ follow such patterns; E2-initial /t/ always voices if E1's last obstruent is homorganic /s/ or /t/ but not when it is non-homorganic /k/. We expect similar patterns to be found in surnames with E2 morphemes other than /ta/. Specifically, surnames with /k/-initial E2 should undergo rendaku when E1's last obstruent is also /k/, as in (75c). (See Section 3.3.2 for details.)

Identity Avoidance will also inhibit rendaku in surnames. A stem-internal restriction on sequences of labial sounds apply to the whole compound surname. When E1's last consonant is /m/ and E2's initial consonant is /h/, rendaku will be blocked since its application would create a labial-labial sequence, as shown in (76). (See Section 3.3.3.)

(76) A would-be labial-labial sequence blocks rendaku

a.	u̲me	+	hara	→	u̲me-hara	*u̲me-**b**ara	'plum-field'
b.	i̲ma	+	hasi	→	i̲ma-hasi	*i̲ma-**b**asi	'present-bridge'

The liquid /r/ in E1 is also expected to inhibit rendaku. I have proposed that stems containing a sequence of /r/ and a voiced obstruent (represented here as rVD) are relatively rare due to a combination of distributional restrictions on the two classes of sounds. Rendaku is thus blocked in surnames with E1-/r/ in order to avoid the creation of a disfavored rVD sequence, as shown in (77). (See Section 3.3.7 for details.)

(77) /r/ in E1 blocks rendaku due to *rVD

a.	hi̲ra	+	ta	→	hi̲ra-ta	*hi̲ra-**d**a	'flat-paddy'
b.	hi̲ro	+	kawa	→	hi̲ro-kawa	*hi̲ro-**g**awa	'large-river'

Other predictions are summarized as follows. Although the nasals /m/ and /n/ in E1 may trigger voicing in surnames with certain E2 such as /ta/ (Sugito 1965), they will not generally exert strong effects on the applicability of rendaku. Following the description by Zamma (2005), I have also proposed that E1-/m/ triggers rendaku less than E1-/n/ because of an indirect influence of historical sound change. (See Section 3.3.5 for details.) /w/ and /y/ in E1 are considered to be neutral consonants and they will not affect rendaku application in surnames. (See Section 3.3.6.)

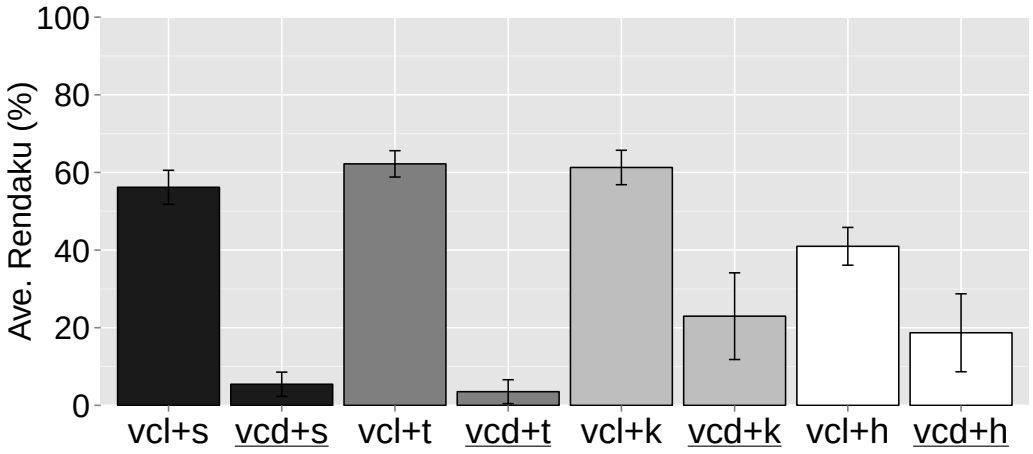


Figure 4.1: Average rendaku rates by E1-obstruent voicing and E2-obstruent place (corpus)

4.2.4 Results

4.2.4.1 Strong Lyman's Law and Identity Avoidance effects

Let us first examine the effects of Strong Lyman's Law and Identity Avoidance.[14] Figure 4.1 plots the average rendaku application rates (%) by surname type based on the voicing of E1's last obstruent (voiceless or voiced, labeled "vcl" or "vcd" respectively) and the place of E2's initial obstruent (/s/, /t/, /k/ or /h/). For example, the bar labeled "vcl+s" includes surnames such as /na̲ka-s̲ima/ 'center-island' and that labeled "vcd+s" includes surnames such as /na̲ga-s̲ima/ 'long-island.' Error bars represent ±1 standard errors. As can be seen, surnames with a voiced obstruent in E1 (labels underlined) generally have lower rendaku rates than those with a voiceless obstruent in E1. This suggests that the presence of a voiced obstruent in E1 inhibits rendaku application in surnames. In other words, Strong Lyman's Law is at work in surnames. (See below for a statistical analysis.)

Let us turn to the effects of Identity Avoidance. Figure 4.2 plots the average rendaku application rates (%) by surname type based on the place of the last obstruent of E1 (/s/, /t/ or /k/) and of the initial obstruent of E2 (/s/, /t/, /k/ or /h/). For example, the bar labeled "s+s" includes surnames

[14]The results shown here are slightly different from those presented in Tanaka (2017). This is because the data in Tanaka (2017) include fewer surnames (*n*=322) and are only based on the readings collected from mixi.

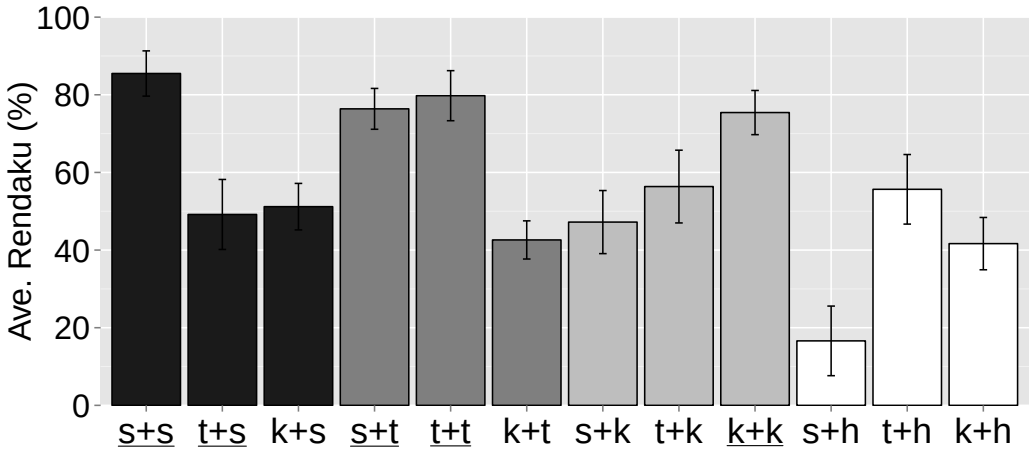


Figure 4.2: Average rendaku rates by E1-E2 voiceless obstruent combinations (corpus)

such as /nis̲i-s̲ima/ 'West-island' and the bar labeled "s+k" includes those such as /is̲i-k̲awa/ 'stone-river.' As can be seen, surnames that have a sequence of homorganic voiceless obstruents (which are underlined and predicted to undergo rendaku) generally have higher rendaku rates than those that do not. Notice that /k+k/ has a higher rendaku rate than /s+k/ and /t+k/, as predicted. The applicability of rendaku also seems to differ among surnames with homorganic obstruent sequences. The rendaku rate of /s+t/ is about as high as that of /t+t/. The rate of /t+s/, on the other hand, is no higher than that of non-homorganic /k+s/. This suggests that, in the case of /s/ and /t/ which share voicing and place but not manner, the order of the sequential obstruents also matters. (See more discussion below.) The hypothesis did not have particular predictions about surnames with /h/-initial E2, but some differences are found among them. The rendaku rate of /s+h/ is particularly lower than those of /t+h/ and /k+h/. (The pattern does not hold in a productivity test, however; see Section 4.3.4.1.)

To test the effects of the boundary-spanning Lyman's Law and Identity Avoidance statistically, a mixed-effects logistic regression model was run using the `glmer()` function of the lmerTest package (Kuznetsova et al. 2013) built on lme4 (Bates et al. 2011) in R (R Development Core Team 1993-2017). To use a binomial logistic regression model, instead of the rendaku rate of each surname, the presence of rendaku voicing (either rendaku or no rendaku) in each occurrence of surnames based on the calculated household numbers was entered into the model as the dependent

variable. For example, the surname UR /naka-ta/ 中田 discussed above is considered to occur with rendaku 39,350 times and without rendaku 4,499 times (see Section 4.2.2). The independent variables included the voicing of E1's last obstruent ("E1-voicing": voiced or voiceless) and the homorganicity of E1's last voiceless obstruent and E2's initial voiceless obstruent ("Homorganicity": whether voiceless obstruents occurring across the E1-E2 boundary share place or not). Since sequential coronal voiceless obstruents may disagree in manner, as in /t+s/ and /s+t/, they were coded as such to be differentiated from those with total identity, as in /s+s/ and /t+t/, and coronal manner mismatch was also entered into the model as an independent variable ("Manner-Mismatch"). According to previous studies, Japanese speakers show stronger dispreferences for sequences of identical moras (e.g. [ka-ka]) than for sequences of moras with identical consonants but different vowels (e.g. [ka-ke]) in nonce-word experiments (see Kawahara and Sano 2014a, 2016). To test the effect of moraic identity, surnames which have a sequence of identical moras underlyingly (e.g. /naka-kawa/ 'center-river') were coded as such to be differentiated from those with just identical consonant sequences (e.g. /take-kawa/ 'bamboo-river'), and moraic identity was included as one of the independent variables ("MoraIdent"). Additionally, the place of E2's initial obstruent ("E2-iniC": /s/, /t/, /k/ or /h/) and its interaction with each of the four factors above (E2-iniC * E1-voicing, E2-iniC * Homorganicity, E2-iniC * MannerMismatch and E2-iniC * MoraicIdentity) was entered into the model. The length of the first element (monomoraic, bimoraic or trimoraic) was also included in the model. This was intended to take out the effects of E1 length which could potentially skew the data of each target condition.[15] (The detailed result of this factor will be reported in Section 4.2.4.3 below.) Surname URs and E2's were included as random intercepts.

A table of the coefficients for each factor predicted by the model is shown in Table 4.1. The *p*-values were estimated by the Markov chain Monte Carlo method (lmerTest package, Kuznetsova et al. 2013). The model shows that the effect of E1-voicing is significant in that the presence of

[15]For example, monomoraic /ta/ 'paddy' appears very often as E1 (and also as E2). Since one-mora E1 generally promotes rendaku (see Section 2.4.3), surnames with /t/ as E1's last consonant may generally have a high rendaku rate regardless of the other factors that are relevant to the discussion here.

	Estimate	Std. Error	z-value	p-value	
(Intercept)	–0.9726	1.5853	–0.613	0.5395	
E1-voicing	–4.0752	1.4314	–2.847	0.0044	**
E2-iniC-k	1.6401	2.1962	0.747	0.4552	
E2-iniC-s	–2.1435	2.4335	–0.881	0.3784	
E2-iniC-t	–2.2076	2.2126	–0.998	0.3184	
Homorganicity	2.3409	1.1497	2.036	0.0417	*
MannerMismatch	0.5145	1.2190	0.422	0.6730	
MoraicIdentity	2.9200	1.9053	1.533	0.1254	
1μ-E1	2.5355	0.5562	4.559	0.0000	***
3μ-E1	5.4893	1.4899	3.684	0.0002	***
E1-voicing:E2-iniC-k	–1.0580	1.9766	–0.535	0.5925	
E1-voicing:E2-iniC-s	–2.5629	1.8763	–1.366	0.1720	
E1-voicing:E2-iniC-t	–3.0952	1.7997	–1.720	0.0855	.
E2-iniC-k:Homorg	0.5212	1.5688	0.332	0.7397	
E2-iniC-s:Homorg	1.2663	1.7070	0.742	0.4582	
E2-iniC-s:MMismatch	–4.1132	1.9318	–2.129	0.0332	*
E2-iniC-k:MoraIdent	2.8201	2.7371	1.030	0.3029	
E2-iniC-s:MoraIdent	7.3651	3.4635	2.127	0.0335	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '					

Table 4.1: Regression model coefficients table; Effects of E1-obstruents (corpus)

a voiced obstruent in E1 lowers the probability of rendaku application (z=–2.847, p<0.01), and its interaction with E2-iniC is not significant.[16] This indicates that Strong Lyman's Law is at play and inhibits rendaku regardless of E2's initial obstruent. The effects of Homorganicity are also significant in that the presence of a sequence of homorganic voiceless obstruents in E1 and E2 raises rendaku applicability (z=2.036, p<0.05), and there is no significant interaction with E2-iniC. This indicates that Identity Avoidance is at work and avoidance of homorganic voiceless obstruents generally triggers rendaku. MannerMismatch does not have a significant effect (z=0.422, p=0.673;

[16]The intercept of the model here refers to the condition where a surname has a voiceless obstruent in E1, E1 is bimoraic and E2'initial consonant is /h/.

with /t/ as E2's initial consonant) but its interaction with E2-iniC does; it significantly lowers the rate of rendaku application when E2's initial consonant is /s/ (z=–2.129, p<0.05). In other words, although rendaku is usually promoted by Identity Avoidance, the effect is nullified if a surname has a sequence of coronal obstruents which disagree in manner and E2's initial consonant is /s/ (that is, when the sequence is /t+s/). The interaction of MoraicIdentity and E2-iniC is significant; MoraicIdentity has no effect when E2's initial consonant is /t/ (z=1.533 p=0.125) or /k/ (z=1.030 p=0.303) but it raises rendaku applicability when E2's initial consonant is /s/ (z=2.127 p<0.05). That is, rendaku is further promoted in surnames such as /nisi-sima/ 'West-island,' which have a sequence of identical moras with /s/ across the E1-E2 boundary. No significant effect of E2-iniC is found, indicating that rendaku application rates are generally not affected by the place of E2's initial obstruent undergoing the voicing change.

I conclude from these results that stem-internal phonological restrictions such as Lyman's Law and Identity Avoidance apply to the whole compound and affect the application of rendaku in the case of a surname.

4.2.4.2 Effects of E1-sonorants

Let us turn to the rendaku patterns of surnames with a sonorant in E1's last syllable. Figure 4.3 shows the average rendaku application rates (%) by surnames with E1-nasals. (Error bars represent ±1 standard errors.) For example, the bar labeled "m+s" includes surnames such as /yama-saki/ 'moutain-cape' and the bar labeled "n+s" includes /kana-saki/ 'gold-cape.' The graph shows that names generally have lower rendaku rates when E1's last consonant is /m/ than when it is /n/, except in the case of those with /h/-initial E2. This suggests that /m/ in E1 triggers rendaku less than /n/. /m+h/ (underlined) is expected to resist voicing since the sequence [m...b] would violate the labial cooccurrence restriction on rendaku application. However, it shows a particularly high rendaku rate, going against the prediction that stem-internal place cooccurrence restrictions affect rendaku, although it should be noted that the condition also has a particuarly small number

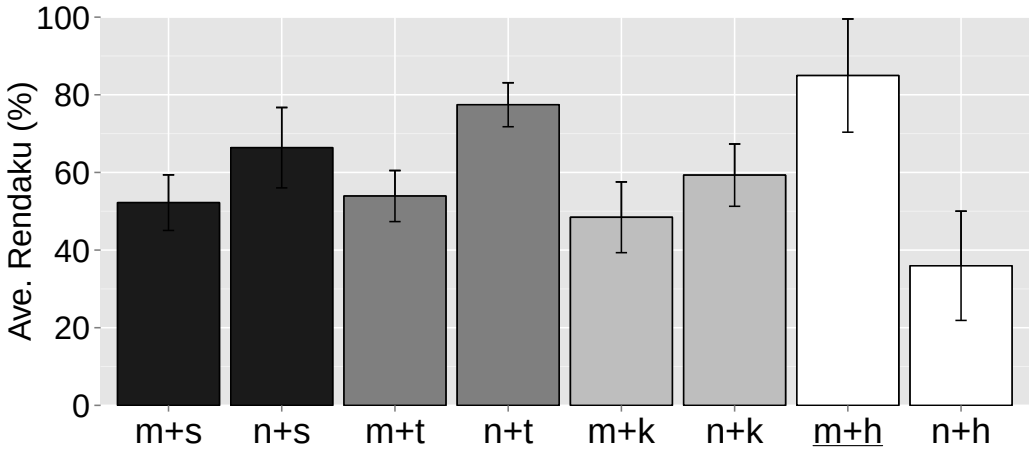


Figure 4.3: Average rendaku rates by E1-consonant voicing and E2-consonant place (corpus)

of samples (*n*=6) compared to the other conditions,[17].

Figure 4.4 shows the average rendaku rates of surnames with E1-approximants. For example, the bar labeled "r+t" includes /nar̲i-t̲a/ 'complete-paddy' and "y+t" includes /toy̲o-t̲a/ 'rich-paddy.' The sample size of surnames with /y/ and /w/ is relatively small[18] and it is not easy to make a solid generalization; but by and large, /y/ and /w/ in E1 do not seem to promote or inhibit voicing in any consistant way. The overall average rendaku rates of surnames with E1-/y/ and those with E1-/w/ are 47.12% and 41.93% respectively. /r/ in E1, on the other hand, shows a somewhat more consistent pattern. It can be seen that the rendaku rates of surnames with E1-/r/ are relatively small regardless of E2's initial consonant, with the overall average being 16.47%. This suggests that /r/ in E1 inhibits rendaku in surnames, as has been suggested by Kubozono (2005).

A mixed-effects logistic regression model was run with rendaku application based on the household numbers of surnames (see above) as the dependent variable. To directly compare the effects of different sonorants on rendaku applicability, the type of sonorant in E1 ("E1-Son": /m/, /n/,

[17] /m+s/: *n*=24; /n+s/: *n*=18; /m+t/: *n*=34; /n+t/: *n*=25; /m+k/: *n*=27; /n+k/: *n*=31; /m+h/: *n*=6; /n+h/: *n*=10.

[18] /r+s/: *n*=39; /y+s/: *n*=13; /w+s/: *n*=13; /r+t/: *n*=29; /y+t/: *n*=9; /w+t/: *n*=9; /r+k/: *n*=35; /y+k/: *n*=14; /w+k/: *n*=16; /r+h/: *n*=12; /y+h/: *n*=3; /w+h/: *n*=11.

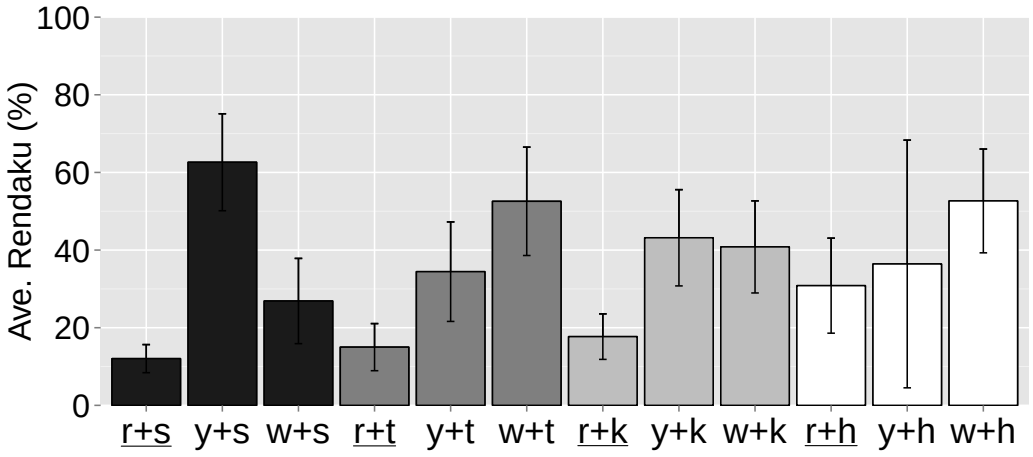


Figure 4.4: Average rendaku rates by E1-approximant and E2-consonant place (corpus)

/r/, /y/ or /w/) was entered into the model as an independent variable. To test the effect of the labial cooccurrence restriction, surnames which would have a sequence of labial consonants by rendaku application (e.g. /ume-hara/ → [ume-**b**ara]) were coded as such and a potential violation of OCP(labial) was included as one of the independent variables ("LabLab"). The other independent variables were the place of E2's initial obstruent ("E2-iniC": /s/, /t/, /k/ or /h/) and the length of the first element (monomoraic, bimoraic or trimoraic). The interaction of E1-Son and E2-iniC was not put into the model this time, because the sample size was generally small and there were too few tokens of the conditions with each combination of E1-sonorant and E2-initial obstruent to reliably test the interaction terms. Surnames and E2's were included as random intercepts.

A table of the coefficients for the model is shown in Table 4.2. The intercept of the model here corresponds to the condition where surnames have /y/ in bimoraic E1 and /h/ as E2's initial obstruent. The presence of /r/ in E1 lowers the probability of rendaku application (z=–3.415, p<0.01). This validates the generalization that /r/ particularly inhibits the voicing alternation in surnames (Kubozono 2005). E1-/m/ has no significant effect (z=1.591, p=0.112) but E1-/n/ raises the probability of rendaku significantly (z=2.599, p<0.01). This can be interpreted as the effects of *NVT and *mVD; E1-nasals generally promote rendaku, but /m/ dampens the effect, causing the difference in the rendaku patterns between surnames with /n/ and those with /m/, as discussed

	Estimate	Std. Error	z-value	p-value	
(Intercept)	–0.05127	2.11753	–0.024	0.9807	
E1-Son-m	1.42652	0.89641	1.591	0.1115	
E1-Son-n	2.28771	0.88029	2.599	0.0094	**
E1-Son-r	–2.94608	0.86271	–3.415	0.0006	***
E1-Son-w	0.11282	1.04296	0.108	0.9139	
LabLab	5.33720	2.43652	2.190	0.0285	*
E2-iniC-k	–0.68749	2.51675	–0.273	0.7847	
E2-iniC-s	–4.83753	2.70452	–1.789	0.0737	.
E2-iniC-t	–6.52579	3.02038	–2.161	0.0307	*
1μ-E1	2.51171	0.78176	3.213	0.0013	**
3μ-E1	4.46941	1.37999	3.239	0.0012	**
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’

Table 4.2: Regression model coefficients table; Effects of E1-sonorants (corpus)

in Section 3.3.5 (also see Zamma 2005). There is no significant effect of /w/ in E1 (z=0.108, p=0.914), which is compatible with the claim that /y/ and /w/ are neutral consonants and do not particularly affect rendaku application. The effect of E2-iniC is also significant. When E2's initial consonant is /t/, the probability of rendaku application is significantly low (z=–2.161, p<0.05). The other E2-initial obstruents /s/, /k/ and /h/ do not have such effects (/s/: z=–1.789, p=0.074; /k/: z=–0.273, p=0.785). LabLab has a significant effect in that it raises the probability of rendaku application. As stated above, this is unexpected since the prediction is that rendaku application should be less likely if it creates a labial-labial sequence within a surname due to the application of the stem-internal labial cooccurrence restriction. It is conceivable that the restriction on /m/ and /b/, which agree in place, does not have as strong effects as the ban on sequences of homorganic voiceless obstruents, which share place and manner. It is surprising, however, that the alternation is even promoted when it creates a labial-labial sequence. Although I have no explanation for this fact, I will come back to the issue when I present experimental results suggesting that the unexpected pattern in real surnames is not actually internalized in Japanese speakers' grammar (see Section 4.3).

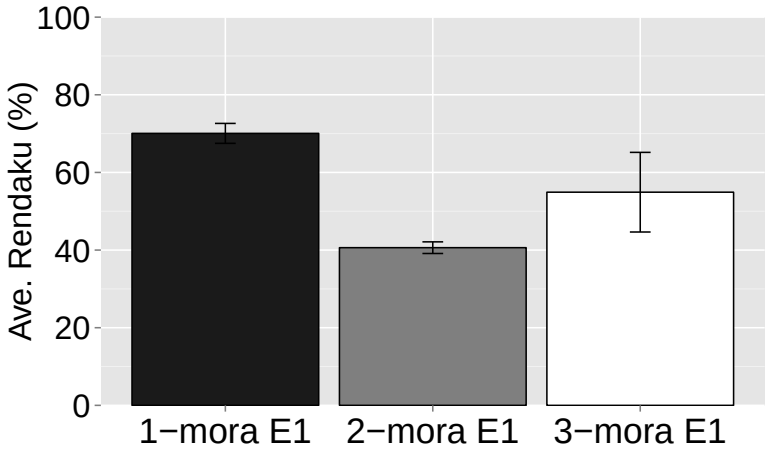


Figure 4.5: Average rendaku rates by E1-length (corpus)

To summarize, the results show that rendaku application in real surnames with sonorants in E1 also follow stem-internal phonology (except the labial cooccurrence restriction).

4.2.4.3 Other patterns

Let us examine other factors that play a role in the application of rendaku in surnames. I will first examine the effect of the length of the first element. Figure 4.5 shows the average rendaku rates of surnames by E1-length.

As can be seen, the rendaku rates of surnames with one-mora E1 and three-mora E1 are generally higher than that of surnames with two-mora E1. The statistical analyses shown above reveal that the rendaku promoting effects of monomoraic first elements are significant both in surnames with an obstruent in E1 (z=4.559, p<0.001) and in surnames with a sonorant in E1 (z=3.213, p<0.01). Similarly, trimoraicity of E1 promotes rendaku significantly in surnames with an obstruent (z=3.684, p<0.001) and those with a sonorant (z=3.239, p<0.01). The results corroborate the observations made in previous studies that the length of E1 affects rendaku in surnames (Zamma 2005) as well as in regular compounds (see Rosen 2001, 2003; Irwin 2016a,b among others).

Let us also examine the effects of so-called "special moras." Compound surnames may have a vowel in the last mora of E1, which is typically the last half of a long vowel, as in /o͟o-sima/ [o͟o-sima] 'big-island,' /i͟i-ta/ [i͟i-da] 'rice-paddy,' or the second half of a vowel-vowel sequence,

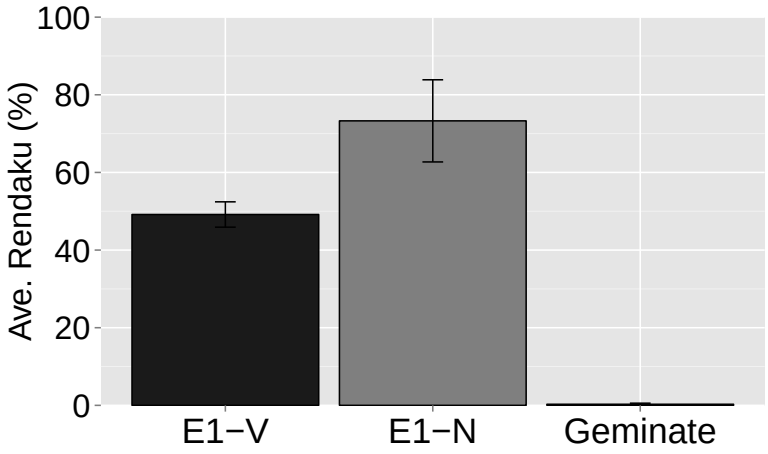


Figure 4.6: Average rendaku rates by special moras in E1 (corpus)

as in /ue-ta/ [ue-da] 'upper-paddy,' /ai-kawa/ [ai-kawa] 'meeting-river,' /mae-hara/ [mae-hara] 'front-field.' Since sequences of a vowel in E1 and a voiceless obstruent in E2 presumably do not incur a violation of any stem-internal phonological restrictions, they are expected not to affect rendaku application. Compound surnames may also have a moraic nasal in E1's last mora, as in /hon-ta/ [hon-da] 'root-paddy' with Sino-Japanese E1 or /kan-saki/ [kan-zaki] 'god-cape' with native E1 undergoing (historical) syncope (derived from original /kami-saki/). In Japanese stems, moraic nasals trigger voicing of following obstruents, which is described under the name of post-nasal voicing (Ito and Mester 1995a,b; see Section 3.3.5). Surnames with a moraic nasal in E1 are thus expected to frequently show rendaku voicing.[19] Although not very frequent, gemination may occur at the E1-E2 boundary of a compound surname, as in /nii-ta/ [nit-ta] 'new-paddy' or as in /kiti-kawa/ [kik-kawa] 'fine-river' through syncope.[20] Given that voiced geminates are generally dispreferred in Japanese phonology (see Ito and Mester 1995b, 1999; Nishimura 2003; Kawahara 2006, 2008, 2011 among others), gemination is expected to block rendaku application (e.g. /nii-ta/ → [nit-ta], *[nid-da]).

Figure 4.6 shows the average rendaku rates of surnames with a vowel in E1 (represented as

[19]It is reported that post-nasal voicing also promotes rendaku in regular compounds. See Irwin (2016a); also see Vance and Asai (2016) for further complications.

[20]The Sino-Japanese root /kiti/ could also be analyzed as /kit/ underlyingly. See Tateishi 1989; Ito and Mester 1996, 2015 for discussion.

"E1-V"), a moraic nasal in E1 (represented as "E1-N") and gemination.[21] As can be seen, surnames with a vowel as E1's last mora have an average rendaku rate of about 50%, suggesting that E1-vowels do not exert any effect on rendaku application. Surnames with a moraic nasal in E1 have a relatively high rendaku rate while those forming a geminate across the E1-E2 boundary show particularly low rendaku rates. The results are expected from the effect of post-nasal voicing and avoidance of voiced geminates. A mixed-effects logistic regression model was constructed with rendaku application as the dependent variable and the type of special mora (a vowel, a moraic nasal, or a geminate), E2's initial consonant (/s/, /t/, /k/ or /h/) and the length of E1 as independent variables. With the intercept of the model corresponding to the condition where a surname has a vowel in bimoraic E1 and E2-initial /h/, it is predicted that the presence of a moraic nasal significantly raises the probability of rendaku application (z=1.966, p<0.05) while gemination lowers the probability (z=–4.451, p<0.001). The effects of E2's initial consonants are not significant. Monomoraicity of E1 significantly raises rendaku applicability (z=3.630, p<0.001). (There are no data with trimoraic E1 and special moras.) That is, the statistical analysis confirms the effects of post-nasal voicing and avoidance of voiced geminates on rendaku application and also indicates that vowels do not affect voicing in any particular way.

4.2.4.4 Summary of the results

To summarize the overall results, the rendaku patterns of real surnames are, for the most part, compatible with the hypothesis that compound surnames are subject to stem-internal phonotactics. Stem-bounded Lyman's Law and Identity Avoidance apply to the whole compound and rendaku applies accordingly. Sonorants in E1 also affect rendaku application so that the surnames will generally conform to stem-phonology; /m/ triggers voicing less than /n/ (which is a historical vestige of the Strong Lyman's Law effects) while /r/ inhibits voicing (due to inferred phonotactics *rVD). The results have also shown two facts which are not predicted by the hypothesis; the asymmetry between /t+s/ and /s+t/ sequences, and the unexpected rendaku promoting effect of

[21] Since these surnames are generally infrequent, data are not found for every single combination with E2's initial consonant. I thus show the overall average rendaku rates.

the labial cooccurrence restriction. I will return to these remaining issues when I present the results of a nonce name experiment in Section 4.3.

4.2.5 A comparison with rendaku in regular compounds

I have shown that stem-internal phonology regulates rendaku application in surnames. In this section, I will give an analysis of the voicing alternation in regular compounds and show that none of the effects discussed above are operative in common nouns. The results will support the claim that the differences in rendaku application between compound surnames and regular compounds arise from the differences in their representations.

To examine the rendaku application patterns in regular compounds, I used "the Rendaku Database v2.5" (Irwin and Miyashita 2013-2016), an extensive database of Japanese compounds with 34,432 entries.[22] Previous studies (Irwin 2014a, 2016a among others) have already conducted detailed analyses of the rendaku patterns using the database. As stated above (see Section 2.3.1), Irwin (2014a, 2016a) shows that Strong Lyman's Law has no effect on the voicing alternation in normal compounds. However, since these studies are not particularly intended to examine the differences between proper nouns and common nouns, additional investigation into the data is required in order to fully address the issues at stake in this dissertation. For example, the database contains not only noun-noun compounds but also combinations of other parts of speech, such as noun-verb compounds, noun-adjective compounds, mimetic compounds and so forth. Since compound surnames all have nouns (including deadjectival and deadverbial nouns) as their second element (which is the compound head), we would need to limit the data to regular compounds of the same type in order to make a fair comparison. Also, previous studies have not tested the role of Identity Avoidance as a rendaku trigger in regular compounds.[23] I have proposed above that

[22]The most recent version available is "the Rendaku Database v3.1" (Irwin et al. 2017).

[23]As discussed above (Section 3.3.4), Irwin (2014a, 2016a) tests the role of Identity Avoidance as a rendaku blocker, and finds null results. Kawahara and Sano (2014b) argue that Identity Avoidance plays a role as a rendaku trigger in a nonce word experiment; however, they are only concerned with cases of moraic identity and do not test the effect of consonantal identity per se (see discussion in Section 3.3.4). Also, given that the rendaku patterns in existing words and non-existing words may show discrepancies (see Irwin 2014a, 2016a; Kawahara and Sano 2016), it is worth

cooccurrence restrictions on sequences of homorganic voiceless obstruents within a stem (Kawahara et al. 2006) apply to compound surnames, which are themselves represented as stems, and trigger the voicing alternation. If the restrictions are bounded to the stem level, it is also predicted that voiceless obstruent sequences across the element boundary will not trigger rendaku in regular compounds. Demonstrating inapplicability of Identity Avoidance as a rendaku trigger in the word class would thus be further evidence for the proposed hypothesis.

From the data of Japanese compounds, proper nouns (which are tagged as such in the database) were excluded. In order to make them comparable to surnames as discussed above, compounds with verbs or adjectives as E2, those that are reduplicated including mimetic compounds and dvandvas were excluded. This resulted in 25,889 regular compounds. The pronunciations of the database entries are taken from two large dictionaries (Shinmura 2008; Watanabe et al. 2008). Each database entry is tagged for the presence or absence of rendaku voicing with "+" and "–" signs based on the reading of the word in each dictionary. An entry is marked "+/–" for rendaku application if dictionaries give both the rendaku form and the non-rendaku form as possible pronunciations of the word. In order to treat rendaku application as a binary variable for the purpose of conducting a statistical analysis (see below), I assigned either "+" or "–" to such variable compounds. Words were deemed to be rendaku undergoers if both or either one of the dictionaries gave "+," or both gave "+/–" for the presence of rendaku voicing; on the other hand, they were deemed to be non-undergoers if both dictionaries gave "–," or one dictionary gave "–" and the other gave "+/–."

First, to test the effects of Strong Lyman's Law and Identity Avoidance, compounds with an obstruent in E1 were analyzed. A mixed-effects logistic regression model was run in the same manner as in the corpus study shown above (see Section 4.2.4). The dependent variable was rendaku application (either rendaku or no rendaku) and the independent variables were the voicing of E1's last obstruent ("E1-voicing": voiced or voiceless), the homorganicity of E1's last voiceless

conducting an independent study with data of real compounds before claiming that Identity Avoidance is really a rendaku promoting factor. Asai's (2014) corpus study also suggests that underlying sequences of homorganic voiceless obstruents in E1 and E2 generally make rendaku more likely to apply; however, Asai's data include both common nouns and proper nouns, and the results are not very informative for our purposes.

obstruent and E2's initial voiceless obstruent ("Homorganicity": whether voiceless obstruents occurring across the E1-E2 boundary share place or not), manner mismatch in an obstruent sequence ("MannerMismatch": /s+t/ and /t+s/ or others), moraic identity ("MoraIdent": whether moras occurring at the E1-E2 boundary are completely identical or not), the place of E2's initial obstruent ("E2-iniC": /s/, /t/, /k/ or /h/) and its interaction with each of the four factors above (E2-iniC * E1-voicing, E2-iniC * Homorganicity, E2-iniC * MannerMismatch and E2-iniC * MoraicIdentity). The length of E1 ("E1-length": monomoraic, bimoraic, trimoraic or longer) was also entered into the model as one of the independent variables. Surnames and E2's were included as random intercepts.

The analysis reveals the following. The intercept of the model here corresponds to the condition where a surname has a voiceless obstruent in E1 which is bimoraic and has /h/ as E2's initial consonant. The presence of a voiced obstruent in E1 slightly lowers the probability of rendaku application but its effect is not statistically significant (z=–1.395, p=0.163). The interaction of E1-voicing and E2-iniC is also not significant (with E2-initial /s/: z=–1.507, p=0.132; E2-initial /t/: z=–1.747, p=0.081; E2-initial /k/: z=–0.398, p=0.691). This shows that Strong Lyman's Law is not operative in regular compounds, replicating the results of Irwin (2014a, 2016a) with a slightly different set of items under analysis. The effect of Homorganicity is also not significant (with E2-initial /h/: z=0.384, p=0.701),[24] nor is its interaction with E2iniC (E2-initial /s/: z=1.208, p=0.227; E2-initial /t/: z=0.148, p=0.882; E2-initial /k/: z=0.961, p=0.337). MoraicIdentity is also not significant (z=1.196, p=0.231). These results indicate that Identity Avoidance (including avoidance of consonantal identity and moraic identity) does not play a role as a rendaku trigger in regular compounds, contrary to what has been suggested from the experimental results in Kawahara and Sano (2014a, 2016).[25] The interaction of MannerMismatch and E2-initial /s/

[24]Unlike in surnames, the data had enough surnames with /h/ in E1's last mora (monomoraic native E1, Sino-Japanese E1 or foreign E1) to test the effect of Identity Avoidance on /h/-/h/ sequences.

[25]Although the reason for the difference between the rendaku patterns in existing compounds and those in nonce words is worth investigating by itself, it is beyond the scope of this dissertation. Speculative explanations could be that, absence of lexical specification, speakers rely more on phonological factors when giving judgments on rendaku in nonce words and/or that their morphological analysis of nonce compounds in experimental settings is not so solid that stem-internal patterns occasionally show up.

was significant in that it lowers the probability of rendaku (z=–2.744, p<0.01). This indicates that the asymmetry between /s+t/ and /t+s/ is also found in regular compounds (even though homorganicity itself does not significantly affect rendaku application). The effect of E1-length was significant; one-mora E1, three-mora E1 and E1 longer than three moras all raise the probability of rendaku application (ps<0.001) as compared to two-mora E1. In summary, the effects of the boundary-spanning Lyman's Law and Identity Avoidance are not found in the rendaku patterns of regular compounds.

The rendaku patterns of regular compounds with sonorants in E1 were also examined. Again, a logistic regression model was run with rendaku application as the dependent variable. The independent variables included the type of sonorant in E1 ("E1-Son": /m/, /n/, /r/, /y/ or /w/), a violation of the labial cooccurrence restriction through rendaku ("LabLab"; E1-/m/ and E2-initial /h/), the place of E2's initial obstruent ("E2-iniC": /s/, /t/, /k/ or /h/), the interaction of E1-Son and E2-iniC, and the length of the first element (monomoraic, bimoraic or trimoraic). Surnames and E2's were included in the model as random intercepts.

The model shows that none of the E1 sonorants exert a significant effect on rendaku application. That is, voicing is not promoted or inhibited by E1-nasals or E1-/r/ in regular compounds, unlike in compound surnames (see Section 4.2.4.2). The effect of E2-iniC is also not significant. Some of the interaction terms between E1-Son and E2-iniC are significant; rendaku is less likely to occur when E1 contains /w/ and E2's initial consonant is /k/ (z=–1.968, p<0.05) or /s/ (z=−2.366, p<0.05). It is interesting that compounds with these sequences in E1 and E2 are less likely to undergo rendaku, but it is not crucially relevant to the hypothesis in question. The effect of LabLab is not significant (z=–0.194, p=0.846). The length of E1 was significant; the probability of rendaku application is higher when E1 is one mora (z=2.330, p<0.05), three moras (z=18.381, p<0.001), and four moras or longer (z=12.826, p<0.001), than when it is two moras. In short, none of the segmental factors of sonorants observed in rendaku in compound surnames is found effective in rendaku in regular compounds.

Overall, the results of the analysis of regular compounds in Irwin and Miyashita (2013-2016) indirectly support the claim that rendaku application in compound surnames is governed by stem-

phonology. As I have shown above, phonological patterns observed within single stems are reflected in the way voicing occurs in compound surnames, which are represented as stems composed of stems. Such stem-internal restrictions do not affect the alternation in regular compounds, since they are not labeled as stems in the grammar.

4.2.6 Summary of the corpus study

In order to test the claim that rendaku application in surnames can be attributed to stem-phonology, I have conducted a corpus study of surnames using social media. An analysis of surnames with an obstruent in the first element reveals that voiced obstruents in E1 generally inhibit the voicing alternation, while sequences of homorganic voiceless obstruents occurring across the E1-E2 boundary promote it. These systematic patterns can be characterized as being derived by the application of stem-bounded Lyman's Law and Identity Avoidance, as I have argued in Chapter 3. The main findings of an analysis of surnames with a sonorant in E1 are the following. /r/ in E1 makes rendaku less likely to occur. Nasals in E1 promote voicing but /m/ dampens the effect. Sequences of /m/ and /h/ raise rendaku applicability. The rendaku promotion by the stem-internal labial cooccurrence restriction was unexpected, but the other patterns are compatible with the claim that rendaku in surnames generally conforms to stem-phonology (either following the synchronic patterns or through diachrony; see Section 3.3). I have also compared the rendaku patterns of surnames to those of regular compounds. The results support the hypothesis that compound surnames and regular compounds are represented differently in the grammar.

4.3 Experiment: Rendaku judgment in non-existing surnames

This section presents the results of a rendaku judgment experiment with nonce surnames as the stimuli. Native speakers of Japanese are presented with surnames composed of non-existing E1 and existing E2, and are asked to judge whether the presented names should undergo voicing or not. The productivity of the rendaku application patterns found in the corpus study will be tested with the experimental results.

4.3.1 The aim of the experiment: Testing productivity

The corpus study presented in the previous section has revealed that rendaku in surnames applies generally in accordance with stem-internal phonotactic restrictions. I have taken this as evidence that compound surnames are represented as stems in the grammar and abide by the phonology of stems. This interpretation rests on the premise that rendaku in surnames is an active phonological phenomenon in the synchronic grammar of Japanese. One may still oppose this assumption by arguing that the sound patterns of proper nouns are all completely lexicalized and that there is no role of phonology in determining whether or not voicing occrus in this word class. This extreme view seems hard to defend, once we consider the fact that Japanese speakers apply rendaku to surnames they have never seen before (see Section 2.6.4) including nonce names (Asai 2014; see below).[26] Nevertheless, it indeed remains to be seen whether they apply rendaku in surnames productively in such a way that is predicted by stem-internal phonotactics.

The data from most previous studies are not suitable to address the issue since they are concerned with the rendaku patterns in existing surnames. One intriguing article is Asai (2014); he conducts a rendaku acceptability judgment task using a stimulus set composed of real and nonce surnames in order to test the effect of Strong Lyman's Law. The results show that participants' acceptability rates are relatively low for surnames with a voiced obstruent in E1.[27] However, Asai's stimuli include a large number of existing surnames, which may have affected the overall results, and the experiment is not meant to test the productivity of the pattern per se.[28] Also, the E2 morpheme of the stimulus surnames is either /s/-initial /saki/ 'cape' or /t/-initial /taki/ 'cascade,' and the effect of Identity Avoidance is not tested systematically. In order to fully address the question of whether the application of rendaku in real surnames governed by stem-internal phono-

[26]Rendaku in regular compounds is also attested in nonce word experiments. See e.g. Vance (1979, 1980b); Ihara and Murata (2006); Kawahara and Sano (2014a,b,c). Also see Kawahara (2016) and references therein.

[27]No statistical analysis is reported.

[28]Asai (2014) uses sixteen surnames in total. They can be categorized into four different types. Frequent existing surnames (e.g. /kana-saki/ 'gold-cape'), infrequent existing surnames (/kasa-saki/ 'umbrella-cape'), non-existing surnames composed of real morphemes (/kaza-saki/ 'wind(y)-cape') and non-existing surnames composed of nonce E1 and real E2 (/kada-saki/ 'NONCE-cape').

tactics is a productive process in Japanese phonology, I conduct an experimental study of the wug test design (Berko 1958), using non-existing surnames composed of nonce E1 and real E2 as the stimuli.

4.3.2 Methods

4.3.2.1 Stimuli

Non-existing surnames composed of nonce E1 and real E2 were used as the experimental stimuli. For first elements, nonsense monomorphemic(-looking) stems of two mora long and of the (C)VCV configuration were created. They can be classified into four main types based on the features of the last consonant; those with a voiceless obstruent, those with a voiced obstruent, those with a nasal and those with an approximant. See the table in (78) which shows examples of the E1 items by the last consonant. The number of items is shown in the column "No."

(78) Nonce E1 items for the creation of non-existing surnames[29]

Last C type	Last C	No.	Examples				
Voiceless	s	14	nesi	yuse	hesa	waso	
obstruent	t	18	yuti	nate	mota	heto	nitu
	k	18	meki	tike	seka	nako	eku
Voiced	z	18	kezi	hoze	yuza	azo	wazu
obstruent	b	18	ubi	sobe	nuba	ebo	tebu
	d	10		ode	heda	mudo	
	g	18	segi	tage	hoga	nego	igu
Nasal	m	15	kemi	time	yuma	hemo	nimu
	n	15	honi	tine	kuna	suno	tonu
Approximant	r	12	niri	mere	yora	saro	
	y	10			kuya	ayo	seyu
	w	4			kewa		

For $(C_1)V_1C_2V_2$ stems, there are twelve consonants that can possibly appear in the C_2 position and five vowels (/i/, /e/, /a/, /o/, /u/) that can appear in the V_2 position. For every possible C_2V_2

[29] A full list of the E1 items is given in Section 4.5.

combination, three to four items were created. For example, I created three /CVti/ words, three /CVte/ words, four /CVta/ words, four /CVto/ words and four /CVtu/ words; this resulted in eighteen words with /t/ as the last consonant (as shown in the column "No."). Not all consonant-vowel combinations were possible. For example, /ru/ is a very common ending of Japanese verbs and it was virtually impossible to create nonsense stems with final /ru/ which would not be homophonous with existing words. /CVsu/ words were also avoided since final /su/ might sound like the verb root /su/ 'do,' making participants infer a complex morphological structure as in /CV-su/. /d/, /y/ and /w/ can only appear in certain vowel contexts and the phonotactically illicit consonant-vowel combinations were not used. /h/ was not used as C_2 of $(C_1)V_1C_2V_2$ stems. This is because the phoneme does not usually appear morpheme-medially (for historical reasons), and its occurrence in the C_2 position would signal a morpheme boundary immediately before, as in /CV-hV/. This might also cause participants to analyze the final /hV/ syllable to be a real monomoraic morpheme; /hi/, /he/, /ha/, /ho/ and /hu/ are all existing words. In total, 170 nonsense E1 stems were created. Additionally, I created another 12 E1 stems (one for each consonant), which would be used as stimuli for the practice session.

Each item was associated with a word definition and an example phrase which would be presented alongside in the task (as will be shown in the Procedure section below). The words' definitions were not very detailed, as in "a type of plant," "a type of terrain," "a word for direction," "a word refering to the size of an object" (i.e. an adjective) and so forth. For example, the word /hesa/ was to be presented as a type of plant, with the example phrase "*Hesa*'s leaves are coloring." The word /moki/ was to be presented as a word refering to a direction, with the example phrase "(I) look at the direction of *moki*." The word /homa/ was to be presented as an adjective with the adjectival suffix /-i/, with the example phrase "I hiked a *homa-i* mountain."

For E2 stems, I used the nineteen morphemes with an initial voiceless obstruent which most commonly appeared as E2 of compound surnames in the corpus data. They include six /s/-initial items: /sawa/ '(mountain) stream, swamp,' /sima/ 'island,' /saki/ 'cape,' /se/ 'riffle,' /sita/ 'bottom,' /saka/ 'slope'; three /t/-initial items: /ta/ 'paddy,' /tuka/ 'mound,' /tani/ 'valley'; six /k/-initial items: /kawa/ 'river,' /kuti/ 'entrance, gate,' /ki/ 'tree,' /kami/ 'top, upper,' /kura/

'warehouse,' /kosi/ 'part across, beyond'; four /h/-initial items: /hara/ '(natural) field,' /hasi/ 'bridge,' /hayasi/ 'woods,' /hata/ '(plowed) field.' A list of the items with kanji is given in (79) below.

(79) Real E2 items for the creation of non-existing surnames

Initial C	No.	Example					
s	6	sawa 沢	sima 島	saki 崎	se 瀬	sita 下	saka 坂
t	3	ta 田	tuka 塚	tani 谷			
k	6	kawa 川	kuti 口	ki 木	kami 上	kura 倉	kosi 越
h	4	hara 原	hasi 橋	hayasi 林	hata 畑		

The E1 items and E2 items were then combined to create non-existing surnames. For example, with E1-/hesa/ and E2-/sima/, a surname /hesa-sima/ 'hesa-island' (with a /s...s/ sequence) was created. With E1-/hoze/ and E2-/ta/, a surname /hoze-ta/ 'hoze-paddy' (with a voiced obstruent in E1) was created. In total, 3230 surnames were created as the stimuli for the main session and 228 surnames were created as the practice session stimuli. The created surnames can be classified based on the type of E1's last consonant and its combination with E2's initial consonant. For example, surnames with a voiceless obstruent in E1 can further be classified into those that have an homorganic sequence underlyingly (e.g. /he͟sa-͟sima/) and those that do not (e.g. /e͟ku-͟sima/). Surnames with a voiced obstruent in E1 are characterized as would-be OCP(voice) violaters with rendaku application (e.g. /ho͟ze-͟ta/ → [ho͟ze-**d**a]). Surnames with /m/ in E1 can also be classified into those with a would-be labial-labial sequence (e.g. /ke͟mi-͟hara/ → [ke͟mi-**b**ara]) and those without (e.g. /ti͟me-sima/). The classifications are illustrated with examples in (80) below.

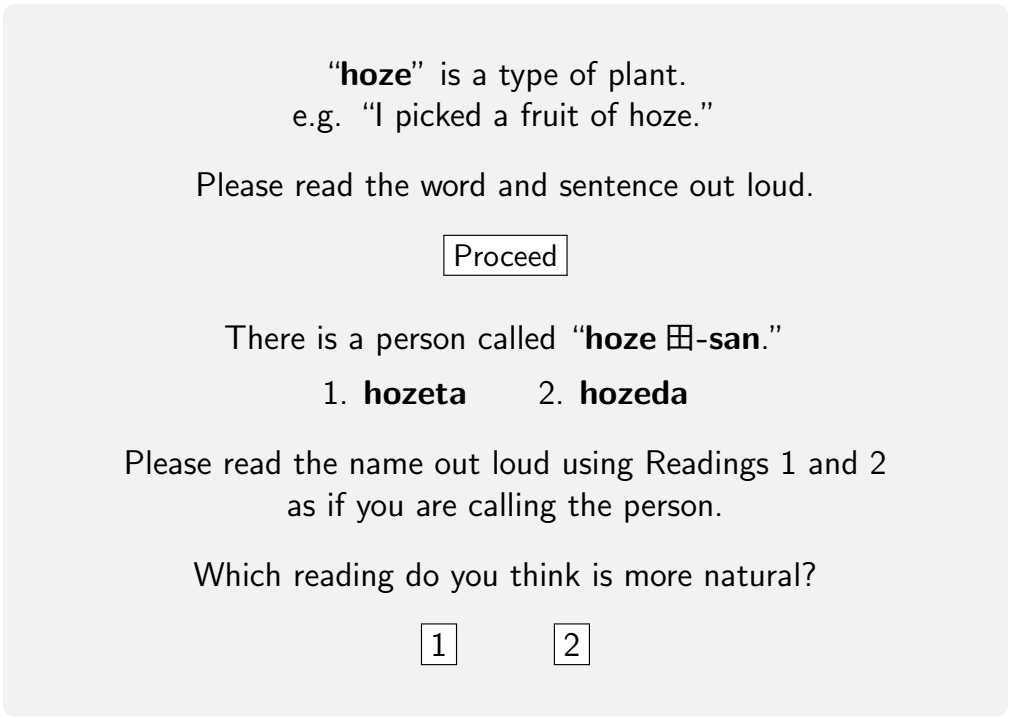
(80) Classification of non-existing surnames

Class			Examples		
E1-voiceless obstruent	s, t, k	Homorganic seq.	hesa-sima yuti-sima	yuti-ta hesa-ta	nako-kawa
		Non-homorg. seq.	eku-sima hesa-kawa	tike-ta yuti-hara	
E1-Voiced obstruent	z, b, d, g		ubi-sima heda-kawa	hoze-ta igu-hara	
E1-nasal	m		time-sima	yuma-ta	hemo-kawa
		Labial seq.	kemi-hara		
	n		kuna-sima	tine-ta	honi-kawa
E1-approx.	r		niri-sima	saro-ta	yora-kawa
	w, y		kuya-sima	kewa-ta	ayo-kawa

4.3.2.2 Procedure

The experiment was designed and run on the Internet-based experiment platform Experigen (Becker and Jonathan 2013). Participants were asked to go to the experiment webpage by clicking a link sent to them by an email. The first page showed a consent form. After agreeing to take the experiment, participants were directed to a general instruction page where they were instructed about the task of the experiment. They were told that they would be presented with uncommon Japanese surnames and be asked to answer questions about their readings. They were also told that some parts of the surnames would be obsolete words or words found only in some regional dialects and they might sound unfamiliar. They were then asked to complete a practice session which had two surname judgment trials. After the practice, they moved on to the main session. An image of the task is shown in (81). The text is translated from Japanese into English.

(81) An image of the experiment task (translated)



At each trial, participants were first presented with an E1 morpheme written in the hiragana script along with its definition and an example sentence, and were asked to read them out loud. On clicking on the Proceed button, new text would appear. A surname composed of the previously presented morpheme as E1 and an existing morpheme as E2 was presented with the honorific suffix /-san/ attached (e.g. /hoze 田-san/). E1 was written in hiragana (e.g. ほぜ /hoze/) and E2 was written in kanji (e.g. 田 /ta/ 'paddy'), which would not indicate rendaku voicing. Participants were given two options for the reading of the surname: one reading with rendaku voicing (e.g. *hozeda*) and the other without (e.g. *hozeta*). They were asked to read the name out loud using both of the reading options and to judge which would sound more natural. To make their response, they clicked on a button marked "1" or "2" according to the number of their selection. Once the response was made, another Proceed button would appear. On clicking on the button, they were taken to the next trial.

In the main session, each participant completed ninety-six judgment trials, each with a different surname. It was designed so that they would each receive forty-eight surnames with a voiceless

obstruent in E1 (/s/, /t/, /k/; forming either an identical sequence or a non-identical sequence with E2's initial consonant), sixteen surnames with a voiced obstruent in E1 (/z/, /b/, /d/, /g/), sixteen surnames with a nasal in E1 (/m/, /n/) and another sixteen with an approximant in E1 (/r/, /y/, /w/). The stimuli were also balanced based on E2's initial consonant; each participant saw twenty-four surnames with /s/-initial E2, eighteen surnames with /t/-initial E2, thirty surnames with /k/-initial E2 and twenty-four surnames with /h/-initial E2. (The numbers are not equally balanced since some consonants have more morphemes than others; see the list of the E2 items in (79) above.) Additionally, some E2 morphemes were set to appear more often than others so that the frequency of E2 would (roughly) match the actual frequency in the corpus. For example, for /s/-initial E2, frequent /sawa/, /sima/ and /saki/ were each presented six times while infrequent /se/, /sita/ and /saka/ were presented two times each. For /t/-initial E2, /ta/ appeared twelve times while /tani/ and /tuka/ each appeared three times. /kawa/ appeared twelve times, /kuti/ and /ki/ each appeared six times, and /kami/, /kura/ and /kosi/ each appeared two times. /hara/ was presented twelve times and /hasi/ was presented six times, while /hayasi/ and /hata/ were each presented three times.[30] For each participant, a distinct set of ninety-six surnames meeting these balancing criteria was created using items randomly selected from the stimulus pool. The stimulus items were presented in a randomized order. Although the same E1 and the same E2 may have appeared multiple times, no participant saw the exact same surname (i.e. the same E1-E2 combination) twice. The presentation order of the rendaku reading and the non-rendaku reading was also shuffled for each trial. For example, the rendaku reading of a surname was presented as Reading 1 in one trial and as Reading 2 in another. (See the task image above.)

4.3.2.3 Participants

Thirty-eight native speakers of Japanese participated in the study. They were recruited through recruitment emails. Besides being a native Japanese speaker, no particular language background was required for participation and a speaker of any Japanese dialect could take the experiment.

[30]The cross-balancing method and randomization made it impossible to present the exact same number of each E2 morpheme to every participant. The numbers are thus approximations of what one participant received.

This is because the results of the experiment were to be compared to the results of the corpus study, which had gathered data of the readings of surnames without taking regional differences into account (see Section 4.2.2). Since the task was orthography-based and no audio stimuli were used, the experimental design did not give participants biases for any particular dialect or for any particular accent patterns which could have affected their rendaku judgments.

At the end of the experiment, participants were asked to answer optional questions about their age, their home region and what dialect(s) had mainly influenced their speech. The responses are summarized in (82) below. (Only the age range and dialects are shown.) Since several participants gave multiple answers as their principal dialects, the number of dialects does not match the total number of participants. (For example, three participants said the Kansai (Osaka/Kobe/Kyoto) dialect had made some influence on their language, but they were all from Tokyo and said Tokyo Japanese was the most influential.) Those who answered "Standard Japanese," including two with the response "Standard Japanese (Chiba)," were counted as Tokyo Japanese speakers. Some dialects were grouped together as being generally in the same dialectal group. (Suruga Japanese and Enshu Japanese were taken together as the Shizuoka dialect.)

(82) Age groups and dialects of experiment participants[31]

Age	No.	Dialect	No.
18–19	4	Tokyo	28
20–24	7	Osaka/Kobe/Kyoto	3
25–29	1	Nagoya/Mikawa	3
30–34	11	Fukuoka/Kitakyushu	2
35–39	2	Shizuoka	2
40–44	6	Hiroshima/Okayama	2
45–49	3	Kagoshima/Kumamoto	1
50–54	3	Nagano	1
55–59	1		

They received an on-line Amazon Japan gift card equivalent of 500 Japanese Yen (about 5 US

[31]The reason why the number of Tokyo Japanese speakers is predominant is most probably that the recruitment was conducted mainly through email lists of universities in Tokyo.

dollars) as a reward for participation.

4.3.3 Predictions

I claim that rendaku application in surnames is productive and that the patterns we have seen in the corpus of existing surnames are internalized in Japanese speakers' minds. The claim makes the prediction that the results of the experiment will be similar to those of the corpus study. More precisely, the effects of Strong Lyman's Law, Identity Avoidance and other patterns driven by E1's last consonants will be attested in Japanese speakers' judgments of rendaku application in nonce surnames.

4.3.4 Results

4.3.4.1 Strong Lyman's Law and Identity Avoidance effects

Let us examine the effects of Strong Lyman's Law and Identity Avoidance. Figure 4.7 plots the average rendaku application rates (%) of surnames with a voiceless ("vcl") or voiced ("vcd") obstruent in E1. As in the graphs presented above, the bar labeled "vcl+s" includes nonce surnames such as /eku-sima/ and that labeled "vcd+s" includes those such as /ubi-sima/. Error bars represent ±1 standard errors. Generally speaking, the presence of a voiced obstruent in E1 (as shown in the underlined conditions) makes the rendaku rate low when compared to a voiceless obstruent in E1, suggesting that Strong Lyman's Law is also operative in rendaku application in non-existing surnames. (See below for a statistical analysis.) Another point to note is that the effects of Strong Lyman's Law do not seem to be as strong as in existing surnames. Although the rendaku rates of surnames with a voiced obstruent in E1 are relatively low, they still moderately undergo rendaku, unlike in real surnames (Figure 4.1; also see Section 4.3.5 for a more detailed comparison of the data.)

Figure 4.8 graphs the rendaku rates of surnames with a voiceless obstruent in E1, sorted by a combination of E1's last obstruent and E2's initial obstruent. The bar labeled "s+s" includes nonce surnames such as /hesa-sima/ and the bar labeled "s+k" includes those such as /hesa-kawa/. The

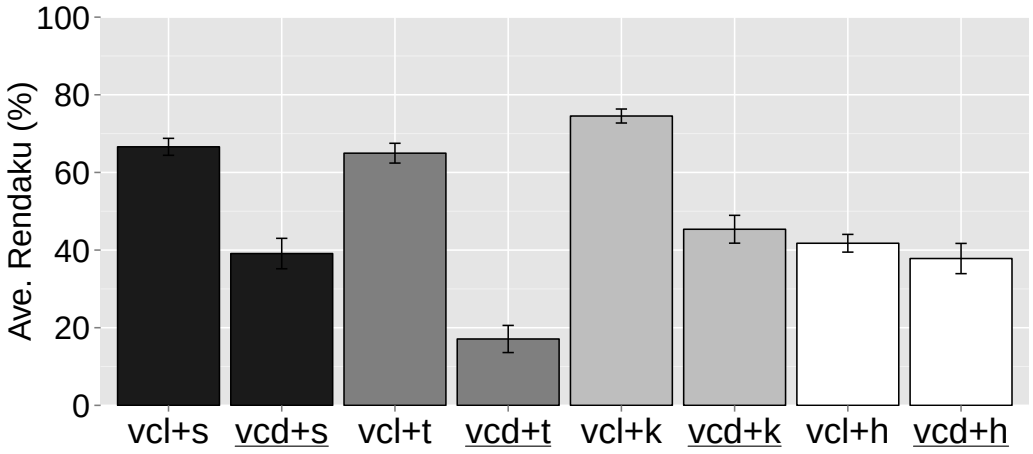


Figure 4.7: Average rendaku rates by E1-C voicing and E2-C place (experiment)

graph shows that the rendaku rates of surnames with a homorganic obstruent sequence (underlined) are relatively high, suggesting that Identity Avoidance affects rendaku application in non-existing surnames as well. Notice that the rendaku rate of /t+s/ is lower than that of /s+t/, showing the same kind of asymmetry as in real surnames. Although some differences in rendaku rates were found among surnames with /h/-initial E2, such a pattern is not seen in rendaku in nonce surnames. Interestingly, the rendaku rates of surnames with no Identity Avoidance violations, such as /k+s/, /k+t/, /s+k/ and /t+k/, are somewhat high compared to what we have seen in existing surnames (Figure 4.2; also see Section 4.3.5 for a detailed comparison between the experimental results and the corpus results), even though they are still lower than their counterparts with Identity Avoidance violations.

A mixed-effects logistic regression model was constructed in a similar way as in the analysis of the corpus results shown above. The dependent variable was rendaku application in participants' responses (either rendaku or no rendaku). The independent variables were the voicing of E1's last obstruent ("E1-voicing": voiced or voiceless), the homorganicity of E1's last voiceless obstruent and E2's initial voiceless obstruent ("Homorganicity": whether voiceless obstruents occurring across the E1-E2 boundary share place or not), manner mismatch in an obstruent sequence ("MannerMismatch": /s+t/ and /t+s/ or others), moraic identity ("MoraIdent": whether moras occurring at the E1-E2 boundary are completely identical or not), the place of E2's initial obstruent

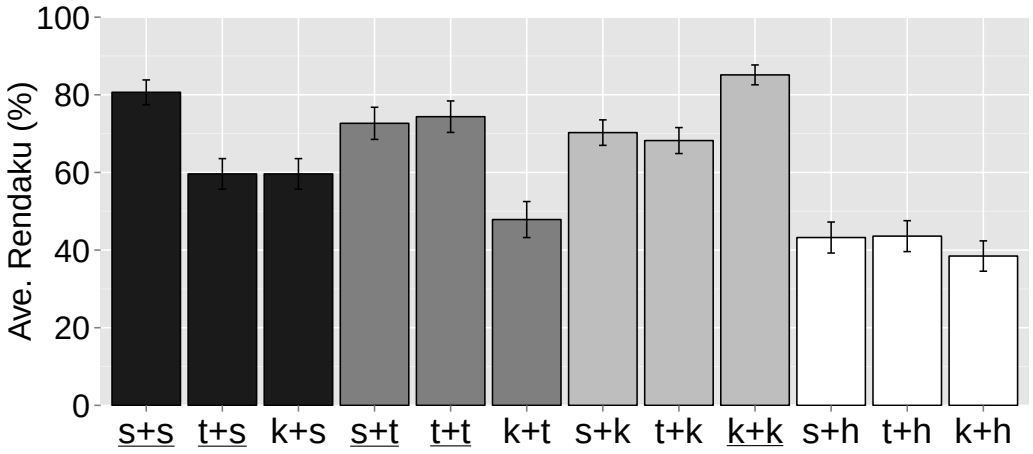


Figure 4.8: Average rendaku rates by E1-E2 consonant place (experiment)

("E2-iniC": /s/, /t/, /k/ or /h/) and its interaction with each of the four factors above (E2-iniC * E1-voicing, E2-iniC * Homorganicity, E2-iniC * MannerMismatch and E2-iniC * MoraicIdentity). The length of E1 was not included in the model since all E1 items were bimoraic. Surnames and E2's were included as random intercepts.

A table of the coefficients of the regression model is shown in Table 4.3. The *p*-values were estimated by the Markov chain Monte Carlo method (lmerTest package, Kuznetsova et al. 2013). The model reveals the following effects. E1-voicing is significant with interactions with E2-iniC. When E2's initial consonant is /h/, there is a trend that the presence of a voiced obstruent in E1 lowers the probability of rendaku application, but its effect is not statistically significant (z=–1.655, p=0.098). However, the effects of E1-voiced obstruents are significant when E2's initial consonant is /s/ (z=–2.259, p<0.05), /t/ (z=–3.465, p<0.001), or /k/ (z=–3.069, p<0.01). That is to say, Strong Lyman's Law generally blocks rendaku in nonce compound surnames, except when E2's initial obstruent is /h/. (See Section 4.3.5 for discussion.) Also, the effect of Homorganicity is significant and no interaction with E2-iniC is significant; a sequence of homorganic voiceless obstruents across the E1-E2 boundary significantly raises the probability of rendaku application regardless of E2's initial consonant (z=2.863, p<0.01). This indicates that Identity Avoidance is operative as a rendaku trigger in nonce surnames. The effect of MoraicIdentity is also significant with no interaction with E2-iniC; a sequence of identical moras across the E1-E2 boundary further

	Estimate	Std. Error	z-value	p-value	
(Intercept)	0.5567	0.8780	0.634	0.5260	
E1-voicing	–0.3997	0.2415	–1.655	0.0980	.
E2-iniC-k	0.6111	1.1228	0.544	0.5863	
E2-iniC-s	–0.7432	1.1349	–0.655	0.5126	
E2-iniC-t	–0.8805	1.3338	–0.660	0.5091	
Homorganicity	1.0641	0.3716	2.863	0.0042	**
MannerMismatch	0.4525	0.3888	1.164	0.2446	
MoraicIdentity	2.3034	0.7814	2.948	0.0032	**
E1-voicing:E2-iniC-k	–1.0171	0.3314	–3.069	0.0022	**
E1-voicing:E2-iniC-s	–0.8487	0.3757	–2.259	0.0239	*
E1-voicing:E2-iniC-t	–1.5124	0.4365	–3.465	0.0005	***
E2-iniC-k:Homorg	–0.2651	0.4616	–0.574	0.5658	
E2-iniC-s:Homorg	0.1375	0.5179	0.266	0.7906	
E2-iniC-s:MMismatch	–1.7636	0.5306	–3.324	0.0009	***
E2-iniC-k:MoraIdent	12.5446	35.5024	0.353	0.7238	
E2-iniC-s:MoraIdent	–1.2495	1.0573	–1.182	0.2373	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’					

Table 4.3: Regression model coefficients table; Effects of E1-obstruents (experiment)

raises the rendaku rate (z=2.948, p<0.01), suggesting that avoidance of moraic identity has even stronger effects than avoidance of consonantal identity (see Kawahara and Sano 2016). The interaction of MannerMismatch and E2-iniC is significant in that coronal voiceless obstruents with manner disagreement lowers rendaku applicability when E2's initial consonant is /s/ (or when the sequence is /t+s/). That is, the asymmetry between /t+s/ and /s+t/ is again found to be a robust pattern. Any other effects in the model such as E2-iniC are not significant.

To summarize the results so far, Strong Lyman's Law and Identity Avoidance both affect rendaku application in nonce compound surnames. This supports the hypothesis that rendaku application in surnames based on stem-phonology is a productive phonological process.

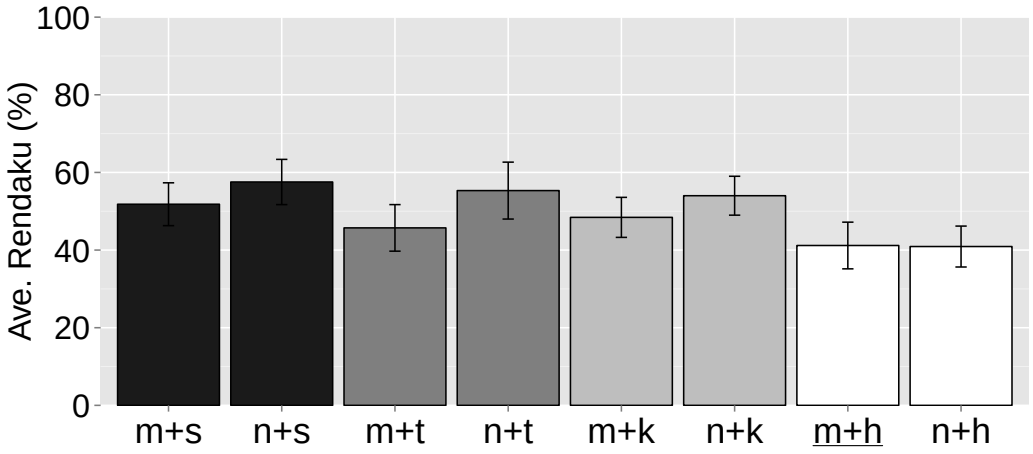


Figure 4.9: Average rendaku rates by E1-nasal (experiment)

4.3.4.2 The effects of E1-sonorants

Let us examine the rendaku patterns of nonce surnames with sonorants in E1. I used the same statistical method as in the analysis of rendaku in real compounds with E1-sonorants discussed in Section 4.2.4.2, except that the model here had participants' judgments as the dependent variable and also included the interaction of the type of E1's sonorant ("E1-Son": /m/, /n/, /r/, /y/ or /w/) and E2's initial conosonant ("E2-iniC": /s/, /t/, /k/, or /h/) because of the larger sample size. Figure 4.9 shows the average rendaku rates of surnames with E1-nasals. Although it looks as if the rendaku rates of surnames with E1-/m/ are generally lower than those of surnames with E1-/n/, the differences are not statistically significant. Neither /m/ or /n/ in E1 affects the probability of rendaku (p=0.581, p=0.485, respectively) and no significant interactions with E2's initial conosonants are found. This indicates that nasals do not exert any effects on the voicing alternation in nonce surnames, unlike in real surnames. Notice that the rate of /m+h/, which would violate the labial cooccurrence restriction through rendaku, is no higher than that of /m+n/. Statistically, the effect of LabLab is not significant (z=–0.979, p=0.327). That is to say, although a violation of the labial cooccurrence restriction still does not inhibit rendaku as is predicted by the hypothesis, its unexpected rendaku promotion effect we have seen in the real name data is not found in nonce surnames. (See Section 4.3.5 for more discussion.)

Figure 4.10 shows the average rendaku rates of surnames with E1-approximants. It is difficult

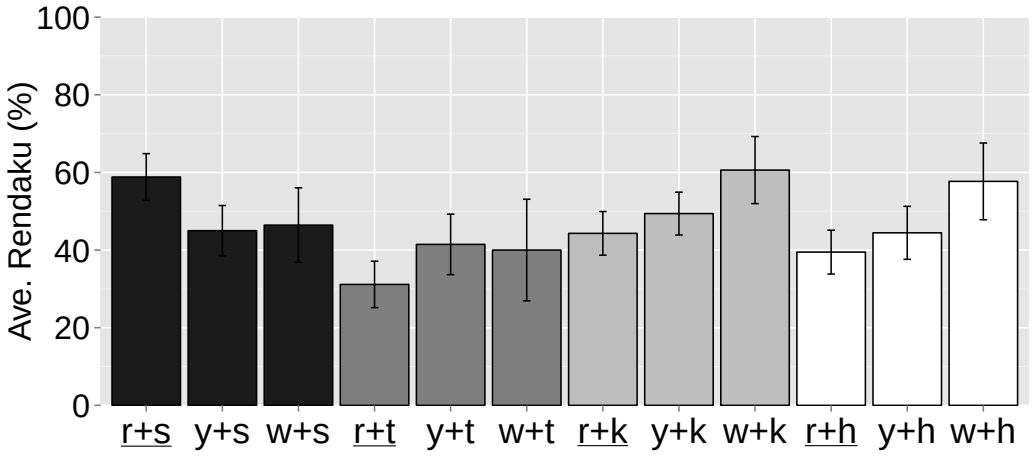


Figure 4.10: Average rendaku rates by E1-approximant (experiment)

to generalize patterns from the graph. The statistical analysis indeed shows that none of the approximant types (/r/, /y/, or /w/) in E1 affects rendaku application in nonce surnames (p=0.491, p=0.697, p=0.271, respectively). None of the interaction terms are significant either. This indicates that the rendaku blocking effect of /r/ in E1, which is operative in real surnames, is not found in nonce surnames.

In summary, unlike nonce surnames with obstruents in E1, those with sonorants in E1 show somewhat different rendaku patterns from their real name counterparts. The segmental effects which are significant in real surnames are not found to be robust factors in rendaku application in nonce surnames. In the next section, I will discuss how the similarities and the differences between nonce surnames and real surnames have arisen, comparing the results of the corpus study and those of the experiment.

4.3.5 Discussion: A comparison to the corpus data

In the previous sections, I have presented a nonce name experiment testing the productivity of rendaku application in compound surnames. The results shed light on two aspects of the phenomenon. On the one hand, Japanese speakers apply the voicing alternation to non-existing surnames in a very similar way as in existing surnames. On the other hand, the rendaku patterns of nonce surnames are somewhat different from those of real surnames. In this section, I will compare

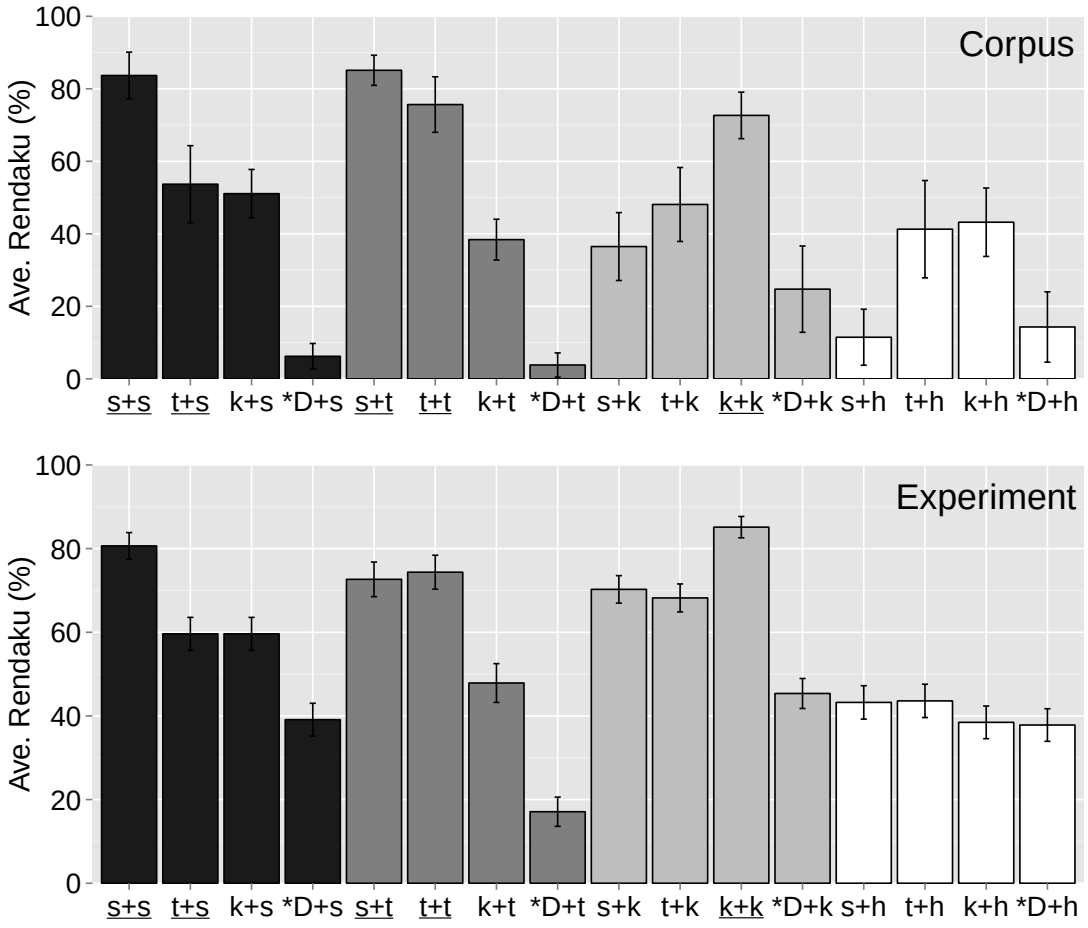


Figure 4.11: Average rendaku rates of surnames with E1-obstruents: Corpus vs. experiment

the experimental results to the corpus study results, and discuss the reasons for the similarities and the differences between them.

Figure 4.11 plots the average rendaku rates of surnames with an obstruent (a voiceless obstruent such as /s/, /t/, /k/, or a voiced obstruent, which is represented by captial /D/ in the label) in the first element altogether. The figure on top represents the results of the corpus study (real surnames) while the bottom figure represents the results of the experiment (nonce surnames). In the corpus results here, I only include the data of surnames with the nineteen E2 morphemes used in the experiment, so that the two graphs can be directly compared to each other. For this reason, the graph of the corpus study looks slightly different from what has been presented earlier (cf. Figure 4.1 and Figure 4.2), but the general patterns are still the same.

As can be seen, the two graphs are similar in that rendaku application rates are affected by Strong Lyman's Law and Identity Avoidance; surnames with a voiced obstruent in E1 (indicated by asterisks as in */D+s/) have relatively low rendaku rates while those with a sequence of homorganic voiceless obstruents at the E1-E2 boundary (underlined) show high rates. When the results of the two studies are taken together, not only do they support the claim that rendkau in surnames is subject to the phonology of stems, but they also provide evidence that the patterns of the voicing alternation in Japanese surnames is not simply memorized by speakers but is a productive process regulated by the phonological grammar of the language.

Another similarity which seems worth mentioning is the asymmetry between /t+s/ and /s+t/. As we have seen, although Identity Avoidance is generally at work in rendaku in surnames, coronal voiceless obstruents which disagree in manner fail to trigger voicing when they appear specifically in the order of /t/ followed by /s/. Compare the rendaku rates of /t+s/ and /s+t/; in both the corpus results and the experimental results, the former is no higher than non-homorganic /k+s/, while the latter is about as high as identical /t+t/. The current hypothesis based on stem phonology does not offer a good explanation for this asymmetry. Kawahara et al. (2006) claim that root/stem-internal sequences of homorganic voiceless obstruents are underrepresented (see Section 3.3.2) and do not particularly mention the order of the consonants. It could be that /s...t/ sequences are actually rarer than /t...s/ sequences in the lexicon, and thus they are avoided more often through rendaku application. Here, I simply stipulate that the restriction of /s...t/ has stronger voicing effects than that of /t...s/, leaving open the questions of whether the pattern is actually found in the lexicon and why the language disfavors /s...t/ over /t...s/ to begin with, even though they are both homorganic sequences.

We also find some differences between the two studies. For example, in the corpus, voiced obstruents in E1 lower the probability of rendaku application regardless of E2's initial consonant, but it is only a trend when it comes to surnames with /h/-initial E2 in the experiment (z=–1.655, p=0.098; see above). It is unclear whether the inapplicability of Strong Lyman's Law in this particular condition (i.e. /D+h/) is a robust pattern in the data, or if it is simply that there are some experimental artifacts which I am not aware of. Supposing that the law's effect is present

but weaker for /h/-initial E2 morphemes, we could porpose an explanation based on a difficulty of learning due to relatively few data samples. Among E2 morphemes with different initial obstruents in the real name data (corpus), /h/-initial ones are the smallest in size.[32] It could be that the patterns with /h/-initial E2 morphemes are somewhat harder to learn, and they did not appear conspicuously in the behavior of speakers in experimental settings.[33] The explanation is, by all means, a mere conjecture, and I will leave the issue open to be examined.

One other noticeable difference between the two graphs is that rendaku rates are generally higher in the experiment than in the corpus. As stated above, experiment participants still applied rendaku moderately even when it would violate Strong Lyman's Law (see especially */D+s/, */D+k/ and */D+h/). It is not simply the case that the law's effects are weaker in the experiment. Notice that the rendaku rates of those with non-homorganic consonant sequences are also relatively high, even though there is no strong trigger of voicing. Indeed, the overall rendaku application rates in the experiment turn out to be higher than those in the corpus (53.25% vs. 46.88%), suggesting that speakers generally applied voicing more often than what they actually observe in real names. This may be an experimental artifact. Recall that the experiment was desgined as a forced choice task with the options of the rendaku reading and the non-rendaku reading. Presenting participants with the rendaku form of a name repeatedly may have prompted their response for voicing. It would be interesting to see whether the results change in a more open-ended design, where speakers are asked to come up with readings of surnames themselves.

Let us now compare the effects of E1-sonorants in the two studies. The graphs in Figure 4.12 show the average rendaku rates of surnames with a sonorant (/m/, /n/, /r/, /y/, or /w/) in the first element altogether. The figure on top represents the results of the corpus study (real surnames) while the one at the bottom represents those of the experiment (nonce surnames). The graphs

[32]The number of surnames by the initial consonant of E2 are the following; /s/: 118; /t/: 168; /k/: 116; /h/: 94. If the data are limited to surnames with the nineteen most common E2s used in the experiment, the distributions are the followings: /s/: 95; /t/: 123; /k/: 90; /h/: 63.

[33]Another difference we can observe is the /s+h/ condition. Its rendaku rate is relatively low in the corpus but not so in the experiment. This may also be due to the smaller data size. Another possible interpretation is that the pattern (the blocking of voicing of /h/ by /s/) is not grounded on any phonological factors, and such an unnatural pattern in the data are not learned by speakers, as will be discussed more in detail below.

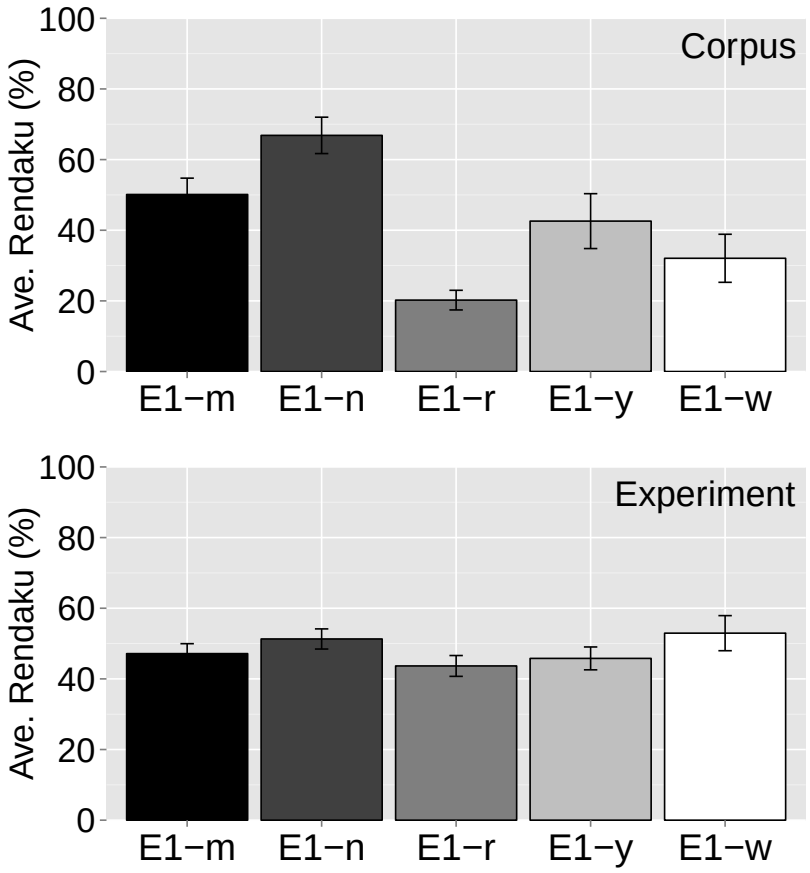


Figure 4.12: Average rendaku rates of surnames with E1-sonorants: Corpus vs. experiment

here do not present the results of the conditions with every E1-E2 consonant combination. As in the analysis of surnames with obstruents in E1, I excluded the data of E2 morphemes which are not used in the experiment, so that the corpus results would be comparable to the experimental results. After the exclusion, the data size turned out to be quite small, with some of the conditions (especially those with /y/ and /w/) having only one or a few samples. For this reason, I collapsed the results of surnames with the same E1-sonorant.[34]

The patterns in the two graphs look somewhat different from each other. As discussed above, E1-/n/ (but not /m/) promotes rendaku while /r/ in E1 inhibits it in the real name data. On the other hand, none of such effects is found in the nonce name data; notice that the rendaku rates in

[34] According to the statistical analysis of the corpus data, the interaction terms of E1-sonorants and E2-initial obstruents are not significant, and thus, the patterns shown here should not differ greatly depending on E2's initial consonant.

the experiment are all close to the chance level, ranging from about 45% to 55% on average. What has caused these differences? What does it mean for a pattern to not show up in a nonce-word experiment while still being real in the data of existing words? I argue that the null findings of the experiment is caused by naturalness biases in learning. More specifically, the rendaku patterns regarding E1-sonorants, which are phonologically unnatural, are not learned by speakers, and thus they are not productive.

As an analysis of the rendaku patterns of real surnames with E1-sonorants, I have posited three Optimality-Theoretic constraints: *NVT, *mVD, and *rVD. In proposing these constraints, however, I have also acknowledged that they are not solidly grounded in phonetics or typological observations. *NVT is intended to explain the behaviors of certain E2 morphemes such as /ta/; unlike the famous post-nasal voicing constraint or *NT which is articulatorily motivated (Hayes and Stivers 1996), it is unclear how the production of a nasal can affect the voicing of a non-adjacent obstruent. *mVD has been proposed to capture a pattern caused by some historical quirk (the sound change of /b/ to /m/), which is specific to Japanese. *rVD is a phonotactic constraint which I have inferred from the distributional skews in the lexicon partly due to history (the development of /r/ in the language); to the best of my knowledge, no such pattern is observed in other languages. A large body of the recent phonological literature suggests that "natural" linguistic patterns are easier to learn than "unnatural" ones (see e.g. Wilson 2006). Following these previous studies, I claim that, even though speakers of present-day Japanese are exposed to these patterns in the lexicon, they have failed to learn them as real phonological patterns.[35] Admittedly, one must be cautious in making an argument based on null results of an experiment; it could be that the design of the experiment was simply not sensitive enough to pick up effects that are real but weak, as I have suggested for the non-effect of Strong Lyman's Law on /h/-initial E2. It is important to note, however, that the very same experiment has proven the productivity of patterns

[35]This is not to say that speakers failed to learn the rendaku patterns of real surnames per se, since they normally have strong intuitions about which names show rendaku or not. The actual patterns are thus learned as lexical patterns, rather than through generalizations based on phonological factors. As will be discussed in more detail below, I argue that it is lexically-specific constraints, as opposed to general phonological constraints, which are involved in the learning of such patterns.

such as Identity Avoidance and OCP(voice), which are claimed to be phonetically motivated and commonly found in languages. This suggests that, unless there are some unknown factors which specifically affected the results of the conditions with sonorants, the effects of the three constraints shown above are at best weak and therefore negligible.

The same argument can be made of the effect of the labial-labial cooccurrence restriction. Although it is not shown in the graphs in Figure 4.12, we have found another difference between the corpus and the experiment. Recall that a /m...h/ sequence unexpectedly raises rendaku rates in the real name data, yielding disfavored [m...**b**], while no such effect is found in the nonce name experiment. (Compare Figure 4.3 and Figure 4.9.) Although it is still unclear why the real name data turned out to be the way they are, I argue that the pattern is unnatural and Japanese speakers have not learned it as part of the rendaku application patterns in surnames. As a result, the effect does not show up in a productivity test, even though it is present in the actual data.

To summarize, the rendaku patterns of nonce surnames revealed by the experiment have similarities and differences when compared to the rendaku patterns of real surnames observed in the corpus. I have argued that the similarities indicate the productivity of the patterns; Japanese speakers apply rendaku based on phonological factors they have inferred from real name data. I have also claimed that the differences are due to a failure in learning; rendaku patterns that are phonologically unnatural are harder to learn, and are not replicated in a nonce word experiment.

4.3.6 Summary of the experiment

Briefly summarizing the findings of the experiment yet again, Strong Lyman's Law and Identity Avoidance have proven to be at work in the rendaku patterns of nonce surnames on the whole. I have taken this as evidence for the productivity of the phenomena as well as evidence supporting the hypothesis that compound surnames are represented as stems. On the other hand, sonorants in E1 do not exert significant effects on rendaku application in nonce surnames. I have interpreted the null results to be evidence that the patterns found in exsiting surnames with sonorants in E1 are not internalized as phonological patterns in Japanese speakers' minds due to naturalness biases in

learning.

4.4 Chapter summary

In this chapter, I have presented the results of a corpus study and an experiment testing the hypothesis proposed in Chapter 3. I collected data of rendaku in existing surnames from social media. The results show that rendaku in surnames generally applies as if it follows stem-internal phonotactic restrictions such as OCP(voice) and Identity Avoidance. This supports the claim that proper names are represented as stems in the grammar despite their compound structures. I also conducted an experiment where Japanese speakers were asked to judge rendaku applicability in nonce surnames. The experimental results show that speakers apply rendaku to nonce names generally in accordance with stem-internal phonotactics just as in real surnames. This further provides evidence for the productivity of rendaku application in surnames governed by stem phonology. I have also compared the rendaku patterns of existing surnames and those of non-existing surnames. I have argued that only natural phonological patterns are internalized as productive processes in the grammar of Japanese speakers.

4.5 Appendix

(83) Nonce E1 items

E1-C	Item								
s	hesa	mesa	nesa	nisa	misi	nesi	yusi	hese	yuse
	mese	hise	keso	naso	waso				
t	kota	ota	mota	meta	heti	meti	yuti	hotu	hetu
	nitu	watu	hete	hite	nate	heto	yuto	muto	meto
k	noka	seka	hika	toka	moki	meki	neki	eku	miku
	heku	neku	soke	tike	eke	moke	nako	yuko	meko
z	yuza	kuza	moza	hezi	mizi	kezi	yuzi	hozu	yazu
	wazu	huze	hoze	yuze	noze	azo	huzo	kezo	nozo
b	yoba	kuba	nuba	hobi	kebi	ubi	ebu	yubu	nubu
	tebu	yube	obe	sobe	tube	ebo	nabo	sabo	ubo
d	heda	yuda	koda	yode	kude	nide	ode	hado	mudo
	modo								
g	hoga	hega	nuga	uga	segi	tagi	yogi	igu	nogu
	tegu	higu	huge	yage	tage	yugo	nego	nogo	sego
m	homa	yuma	kima	homi	nemi	kemi	yumu	nimu	timu
	eme	nime	time	hemo	yumo	semo			
n	hona	kuna	mena	honi	ini	koni	hinu	yonu	tonu
	yune	tine	wane	hano	muno	suno			
r	sera	mora	yora	miri	niri	meri	mere	sire	sere
	aro	haro	saro						
y	kuya	nuya	tiya	keyu	heyu	noyu	seyu	ayo	suyo
	tayo								
w	howa	kewa	nuwa	newa					
Fillers	mose	wato	neka	koze	subo	kode	yogu	kema	huno
	yuro	kuyo	tewa						

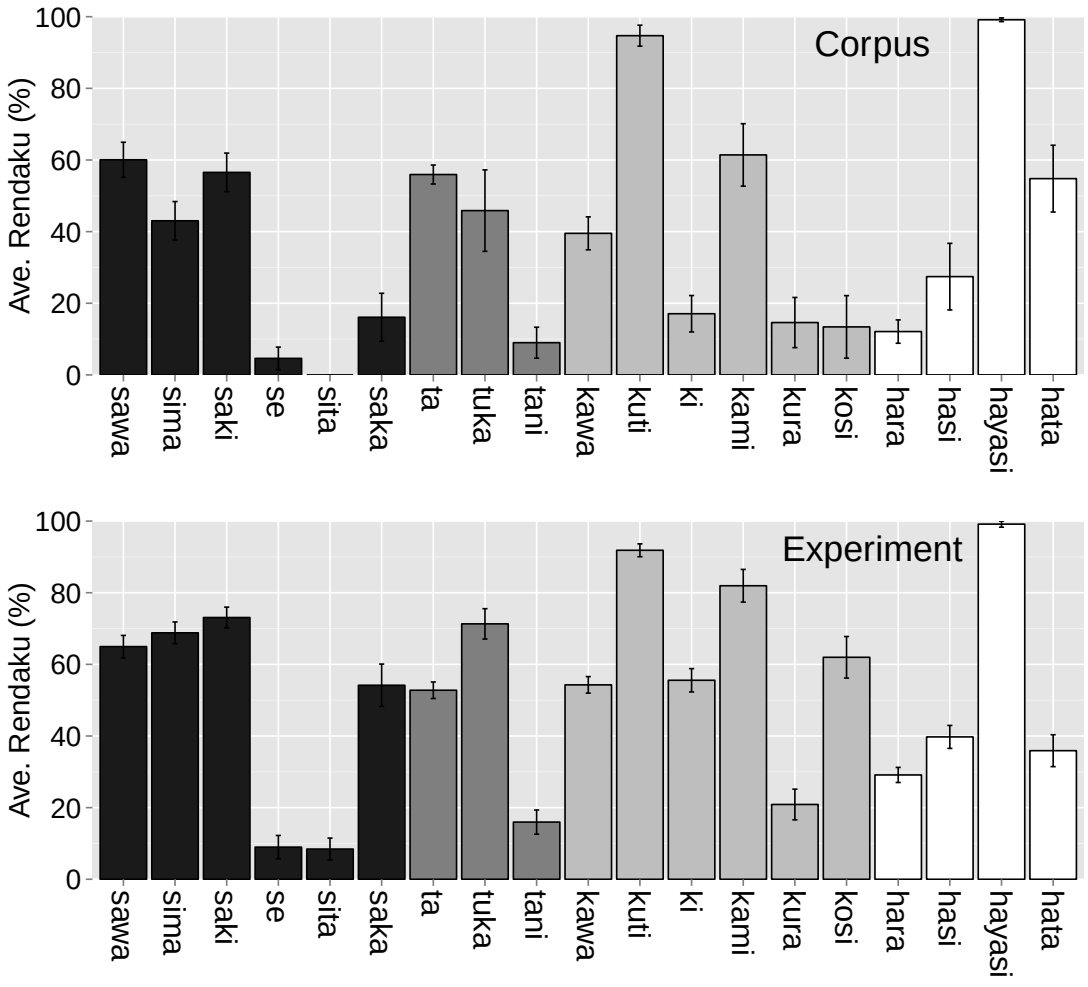


Figure 4.13: Average rendaku rates of surnames by 19 E2's: Corpus vs. experiment

CHAPTER 5

Grammar: A MaxEnt model with lexically-specific constraints

5.1 Chapter overview

In this chapter, I propose a grammar model which aims to account for rendaku application in surnames. The model's predictions will be tested through learning simulations. Section 5.2 first describes the main issues concerning the modeling of the sound patterns of surnames. Section 5.3 presents a probabilistic grammar couched within the framework of Maximum Entropy Harmonic Grammar (see Goldwater and Johnson 2003) along with the constraints and learning biases to be included in the model. Section 5.4 shows the results of learning simulations testing the efficacy of the model. Section 5.5 discusses implications of the results.

5.2 A challenge: Modeling lexicalization and productivity

In the previous chapter, I have shown that Japanese speakers apply rendaku voicing to nonce surnames at rates similar to rendaku application in real surnames, generally in accordance with stem-internal phonological restrictions. This indicates that the voicing alternation in surnames is a productive process and it must be accounted for by phonological grammar. The goal of phonological theory is to construct a model of grammar which can correctly generate not only sound patterns observed in the data of exisisting words but also those of novel words given their input representations. In other words, a good phonological grammar model which accounts for the phenomenon of rendaku in surnames should be able to replicate the behaviors of Japanese speakers who produce the rendaku patterns of nonce surnames as in the experiment, as well as the rendaku patterns of real surnames as they are observed in the corpus study.

Modeling the rendaku application in surnames poses a challenge for current theories of phonology, since the phenomenon involves a lot of variation, and the variability is governed by interactions of both phonological and lexical factors. As stated above, phonological factors play an important role in the productive aspect of the alternation; in Japanese speakers' judgments, non-existing surnames undergo voicing in a parallel manner with existing surnames of similar phonological shapes. On the other hand, the rendaku patterns of real surnames are also highly constrained by lexical factors; the presence of voicing in a given surname must be, to a greater or lesser extent, determined based on the lexical properties of the surname itself. As will be discussed in more detail below, it is this coexistence of phonologically-driven productivity on one hand and lexicalized sound patterns on the other which makes it particularly difficult to model rendaku application in surnames.

As an illustration of the complexity of rendaku in compound surnames, let us take the case of surnames with a /k...k/ sequence occurring across the boundary between the first element (E1) and the second element (E2), as in the existing surname /na̲ka-k̲awa/ 'center-river' or non-existing /na̲ko-k̲awa/ 'nonce-river.' As we have seen, voicing tends to occur in such names in order to dissimilate the sequential homorganic voiceless obstruents (e.g. [na̲ka-**g**awa], [na̲ko-**g**awa]). Voicing by Identity Avoidance is not a hard-and-fast rule since rendaku application is also affected by other (inhibiting and promoting) factors such as a general ban on a change in voicing, namely IDENT(voice) in OT terms, and a general requirement for a compound to undergo voicing, represented as the constraint REALIZE-MORPHEME (see Ito and Mester 1986, 1998, 2003 or Section 3.3.8 for discussion). As a result of the interactions of all these factors, rendaku tends to apply relatively often in both existing and non-existing surnames which contain /k...k/ sequences (with the rendaku rates of around 70% to 85%; see Section 4.3.5).

Probabilistic grammars, including Maximum Entropy Harmonic Grammar (see Goldwater and Johnson 2003) described in more detail below (Section 5.3.1), are well suited for the kind of data illustrated above. Simply put, these stochastic grammar models can infer the significance of each phonological factor from the sound patterns of existing words and predict probabilistically how a given novel word will be pronounced on the basis of its phonological configuration. A grammar

model of this kind would be able to learn from the data of existing surnames that, for example, those containing a /k...k/ sequence often undergo rendaku (e.g. about 80% of the time) and project similar rates of voicing (i.e. about 80%) in non-existing surnames with /k...k/. The process indeed resembles the learning and production of rendaku patterns by Japanese speakers. They have been exposed to existing surnames and come up with generalizations about rendaku application based on phonological factors. They then apply the voicing alternation according to those generalizations when they encounter surnames they have never seen before.

However, estimating the probability of rendaku application purely in terms of phonological factors fails to capture the fact that the alternation is also constrained by the lexical propensities of E2 morphemes. As stated above, some morphemes such as /kuti/ 'entrance,' which may be called "rendaku lovers" in Rosen's (2001) terms, show voicing in most of the surnames they occur in, other morphemes such as /kura/ 'warehouse,' or "rendaku haters," resist voicing in most surnames, and yet some others such as /kawa/ 'river,' or "rendaku waverers" (Irwin 2016a), show rendaku moderately. Because of this, even among surnames with identical /k...k/ sequences, rendaku rates differ greatly depending on the second element (e.g. /sa̲ka-ku̲ti/ [saka-guti] 'slope-entrance': 99.80% rendaku in the corpus; cf. /ta̲ka-ku̲ra/ [taka-kura] 'tall-warehouse': 0.002% rendaku). These E2-dependent rendaku patterns are observed both in the corpus and the experiment (see Figure 4.13 in Section 4.5). Thus, in order to accurately predict the rendaku patterns of existing and non-existing surnames, one must somehow incorporate into the model information on how likely each E2 morpheme is to undergo voicing.

Furthermore, rendaku application in surnames may also be affected by the lexical propensities of compounds themselves. For simplicity, let us limit our data to names with /kawa/, which is a rendaku waverer, as the second element. Existing surnames with E2-/kawa/ and a /k...k/ sequence show rendaku voicing relatively often, with the overall rendaku rate of 70.38%. A probabilistic grammar equipped with relevant phonological constraints would learn the pattern, and predict similarly high rendaku rates in non-existing surnames made of nonce E1 with /k/ in its last syllable and /kawa/ as E2 (e.g. /na̲ko-ka̲wa/). The prediction is on the right track in that it is generally

in accord with the results of the experiment.[1] However, the model would actually fail when it comes to predicting the rendaku patterns of real names. Although the oveall average rendaku rate is 70.38%, each of the existing surnames with a /k...k/ sequence has a different profile with respect to voicing. Many of them show quite fixed patterns, either undergoing voicing most of the time (/naka-kawa/ 'center-river': 99.24%; /huka-kawa/ 'deep-river': 99.99%; /kake-kawa/ 'sheer-river': 99.98%) or only rarely (/seki-kawa/ 'barrier-river': 20.52%); some others show voicing about half of the time (/taki-kawa/ 'cascade-river': 49.12%). Notice that an unelaborated probabilistic grammar would treat all these surnames the same, and assign the same probability of rendaku application (namely, about 70%), wrongly predicting variable rendaku patterns in all of them. Thus, a truly successful model of the phenomenon would need to know how submissive or resistant each surname is to voicing, besides the lexical propensities of E2 morphemes and what phonological factors are in play.

The problem is analogous to the case of stress assignment in English as described by Moore-Cantwell and Pater (2016). Three-syllable or longer English words typically have penultimate stress, as in *banána*, or antepenultimate stress, as in *Cánada*. Both stress patterns are commonly attested in the lexicon, and English speakers use both patterns variably in the production of a nonce word which is phonologically alike in an experimental task (e.g. [bəmǽkə]~[bǽməkə]; Moore-Cantwell 2015).[2] However, the variation in the lexicon and that found in the production of nonce words are not really the same; the latter is token (free) variation while the former is so-called lexical variation. That is, the pronunciation of each of the existing words is usually fixed and does not allow the other stress pattern (e.g. *banána* and never **bánana*; *Cánada* and never **Canáda*). Moore-Cantwell and Pater (2016) point out that a typical probabilistic grammar model would be appropriate for the variable pronunciations of nonce words, but fail to capture the lexically-determined variation in existing words, namely, the fixed stress patterns of individual

[1]The average rendaku rate of nonce surnames with a /k...k/ sequence is 82.05%. As discussed above (see Section 4.3.5), rendaku rates are generally higher in the experiment.

[2]The stress pattern is also affected by the vowel in the final syllable. See Moore-Cantwell (2015) for the details.

items.[3]

Returning to rendaku in Japanese surnames, the issue at stake is similar in essence. On one hand, a grammar model must be able to generate variable rendaku patterns in non-existing surnames, with phonologically-based generalizations learned from the distribution of existing surnames. On the other hand, it must also correctly produce the rendaku patterns of existing surnames, which are highly lexicalized and generally involve less of the token-based kind of variation. In what follows, I extend Moore-Cantwell and Pater's (2016) approach to the issue and propose a probabilistic grammar combined with a set of constraints associated with each E2 morpheme and each compound surname. Implementation of the model further requires the addition of learning biases on the lexically-specific constraints. I will show below that the proposed grammar accounts for both the productive and lexicalized aspects of rendaku application in surnames.

5.3 The grammar model

In this section, I propose a grammar model within the framwork of Maximum Entropy Harmonic Grammar which solves the issue of the coexistence of lexicalization and productivity in rendaku in surnames. I will first give a brief overview of Maximum Entropy models. I will then introduce constraints to be used, including general and lexically-specific constraints. I will also discuss how learning biases on constraints are implemented in the model.

5.3.1 MaxEnt grammar

Maximum Entropy (MaxEnt) models refer to log-linear models which have been used in a wide range of scientific fields. The formalism has been implemented within the basic architecture of the constraint-based grammar known as Harmonic Grammar (Legendre et al. 1990), which is closely related to Optimality Theory (Prince and Smolensky 1993/2004), and it has been applied to prob-

[3]The challenge of modeling both kinds of data has been recognized but not always addressed in the literature (see Moore-Cantwell and Pater 2016 for discussion; also see Zuraw (2000) for a proposal).

lems in phonology (e.g. Goldwater and Johnson 2003; Jäger 2007; Hayes and Wilson 2008; Hayes et al. 2009 to name a few). As is mentioned earlier, a Maximum Entropy Harmonic Grammar, or MaxEnt HG, can learn and generate a probability distribution over a set of phonological forms based on their well-formedness.

A Harmonic Grammar is composed of constraints which are associated with numerical weights, instead of strict rankings as in classical OT. These constraint weights are represented by real numbers and reflect their "strength" or importance in the evaluation of candidates. At the time of evaluation, the "harmony" or penalty score of a candidate is calculated by first multiplying the number of violations of each constraint by the associated weight and then summing such violation scores. Suppose some output y of input x is in evaluation. There are n number of constraints (C_1, C_2, ..., C_n), and constraint C_k has some weight w_k. $C_k(x, y)$ denotes the number of violations assigned by constraint C_k to output y given its input x. The harmony value of the mapping of input x to output candidate y, denoted as $H(x, y)$, can be defined as in (84).

(84) Definition of harmony

$$H(x, y) = \sum_{k=1}^{n} w_k C_k(x, y)$$

Calculation of penalty scores is illustrated in the OT-style tableau in (85). The w shown below each constraint represents the weight of that constraint. The calculated harmony value of each candidate is given in column H. In this study, I posit that constraint weights are all non-negative real numbers and each constraint violation receives a score of minus one, reflecting the fact that it is a penalty.[4]

[4]One can also assume that constraints weights are all non-positive numbers and each constraint violation scores plus one. Another way of implementing the same effect in a Maxent HG model is to negate the penalty score in the process of calculating eHarmony (see below), as is done in some studies (see e.g. Hayes and Wilson 2008). Models with these different calculation methods are notational variants and produce the same results, and I am not particularly committed to any of them.

(85) Tableau showing penalty scores

/x/	C_k w_k=3.0	C_ℓ w_ℓ=2.0	C_m w_m=1.0	H
y	–1			–3.0
z		–1	–1	–3.0
w			–1	–1.0

For instance, candidate z incurs no violation of constraint C_k, one violation of constraint C_ℓ, and one violation of constraint C_m. Its penalty score is thus 3.0*0 + 2.0*–1 + 1.0*–1, or –3.0. In this system, the smaller (i.e. closer to zero) the harmony is, the more well formed an output is. For example, in (85), candidate w is the most well formed among the output candidates for input x. Constraints with higher weights have stronger effects, in that violating those constraints results in bigger penalty scores.

MaxEnt HG further turns penalty scores into probabilities over the set of output candidates. It first calculates the exponential function of the harmony value; that is, it raises e, the base of natural logarithms (about 2.718), to the power of the harmony. I call the resulting figure "eHarmony" (Wilson 2014). Then, the probability of output y mapped from input x, denoted here as $P(y|x)$, is calculated by dividing its eHarmony by Z, which is the sum of the eHarmony values of all possible output candidates for that input x (i.e., all y in the set $Y(x)$). This is defined in the formula in (86).

(86) Formula for calculating the probability of an output candidate

$$P(y|x) = \frac{\exp(\sum_{k=1}^{n} w_k C_k(x,y))}{Z} \quad \text{, where}$$

$$Z = \sum_{y \in Y(x)} \exp(\sum_{k=1}^{n} w_k C_k(x,y))$$

The tableau in (87) illustrates the calculations using the same examples from (85). The eHarmony values and probabilities of candidates are shown in column eH and column P respectively. As can be seen, the grammar predicts that, for input x, candidate w surfaces about 78.70% of the

time while candidate y and candidate z each surface about 10.65% of the time.

(87) Tableau showing the probabilities of candidates

/x/	C_k w_k=3.0	C_ℓ w_ℓ=2.0	C_m w_m=1.0	H	eH	P
y	–1			–3.0	$e^{-3.0}$=0.0498	0.1065
z		–1	–1	–3.0	$e^{-3.0}$=0.0498	0.1065
w			–1	–1.0	$e^{-1.0}$=0.3679	0.7870

As can be seen, the model assigns probabilities to possible output forms, rather than selecting one single winner for each underlying representation. As mentioned above, the grammar of this kind is well suited for modeling phonological phenomena involving free variation, such as rendaku in surnames (e.g. /naka-ta/ → [naka-ta]: 80.4%, [naka-da]: 20.6%).

Another appealing point of using MaxEnt HG is its mathematically-defined learning algorithm. When a MaxEnt model is provided with learning data along with a set of constraints, it finds the constraint weights that maximize the probability of the observed data. The log probability of the data, or $P(D)$, can be calculated by taking the sum of the log probabilities of all pairs of output y and its input x in the data ($\{(y_1|x_1)...(y_n|x_n)\}$), as is defined in (88).

(88) The log probability of the observed data

$$P(D) = \sum_{i=1}^{n} log P(y_i|x_i)$$

The constraint weights learned from the training data can further be used to make predictions about the phonological patterns of novel words. Given an input form and output forms, the model will assign probabilities over the set of output candidates based on their constraint violations. The validity of the model can then be evaluated by testing the predicted probabilities against the data observed in a nonce-word experiment. I thus adopt MaxEnt HG as the grammatical framework in order to model the rendaku patterns in existing and non-existing surnames. I will provide a MaxEnt model with the rendaku patterns of real surnames in the corpus as learning data and let it predict

the rendaku patterns of nonce surnames used in the experiment with the learned constraint weights. In the following sections, I will introduce constraints to be included in the model, and discuss how to implement phonological factors and lexical factors which affect rendaku in surnames.

5.3.2 General constraints

As I have shown in the previous chapters, a number of factors play a role in rendaku application in surnames. In order to account for their rendaku patterns with a MaxEnt HG model, I propose (morpho)phonological constraints in light of the results of the corpus study and the experiment. I call the constraints proposed in this section "general constraints," as opposed to "lexically-specific constraints" which will be introduced in the following section.

First, I include in the model two constraints which generally require and prohibit rendaku application, namely, REALIZE-MORPHEME and IDENT(voice) respecitvely. Their definitions are given in (89).

(89) Constraints on rendaku

i. REALIZE-M(ORPHEME):
Every morpheme in the input has a nonnull phonological exponent in the output (Ito and Mester 2003)

ii. IDENT(voice):
Assign one violation mark for every output segment that differs from its input correspondent in the feature [voice][5]

As discussed in Section 3.3.8, I argue that compound surnames contain the linking morpheme [+voice]$_{\mathfrak{R}}$ in their underlying representations, just like regular compounds (see Ito and Mester 1986, 1998, 2003). REALIZE-M requires a surface realization of [+voice]$_{\mathfrak{R}}$, or rendaku voicing. IDENT(voice), on the other hand, penalizes a change in voicing.

For clarity, I illustrate how the two constraints are violated by the rendaku form and the non-rendaku form of a surname with the tableau in (90). Note that the linear order of the constraints

[5]The definition is taken from McCarthy (2008:66).

is not meant to indicate their ranking relationship. [+voice]$_{\mathfrak{R}}$ is denoted as $\mathfrak{R}$ in the input for simplicity.

(90) Violations of constraints on rendaku

Input	Output	REALIZE-M	IDENT(voice)
/naka-$\mathfrak{R}$-ta/	(a) naka-ta	*	
	(b) naka-da		*

Secondly, I propose the constraints in (91) in order to incorporate into the model the effects of Strong Lyman's Law, which we have seen in the corpus study and the experiment.

(91) Constraints on voiced obstruents

i. OCP(voice)-stem:
Assign one violation mark for every occurrence of multiple voiced obstruents within a stem

ii. *#z:
Assign one violation mark for every [z] in stem-initial position

iii. *#b:
Assign one violation mark for every [b] in stem-initial position

iv. *#d:
Assign one violation mark for every [d] in stem-initial position

v. *#g:
Assign one violation mark for every [g] in stem-initial position

OCP(voice)-stem is a stem-bounded constraint but exerts its effect over the entire word in the case of a compound surname, which is represented as a stem itself (see Section 3.3.1). When the constraint applies to a regular compound, it yields the rendaku-blocking effect known as the application of (normal) Lyman's Law (e.g. /kuro-ka<u>b</u>i/ → [kuro-ka<u>b</u>i], *[kuro-**g**a<u>b</u>i] 'black-mold'). The model also has a family of Markedness constraints banning stem-initial voiced obstruents in general (*#D for short), specified for every manner/place (*#z, *#b, *#d, and *#g). Although the constraints penalize the occurrence of a single voiced obstruent by definition, they may achieve a cumulative effect of rendaku blocking with OCP(voice). This is illustrated in the tableau in (92), which shows the violation profiles of surnames with a voiced obstruent in E1.

(92) Violations of constraints on voiced obstruents

Input	Output	OCP(voice)	*#z	*#b	*#d	*#g
/naga-ℜ-sawa/	(a) naga-sawa					
	(b) naga-zawa	*	*			
/naga-ℜ-hara/	(c) naga-hara					
	(d) naga-bara	*		*		
/naga-ℜ-ta/	(e) naga-ta					
	(f) naga-da	*			*	
/naga-ℜ-kawa/	(g) naga-kawa					
	(h) naga-gawa	*				*

As can be seen, rendaku application in a surname with a voiced obstruent in E1 violates not only OCP(voice) but also one of the constraints of the *#D family. Given the architecture of MaxEnt HG, cumulative interactions of constraints may affect the selection of winning candidates (see e.g. Hayes and Wilson 2008).[6] Recall that the effects of Strong Lyman's Law may differ depending on the initial obstruent of E2. For example, it has been revealed that the presence of a voiced obstruent in E1 does not significantly inhibit rendaku application when E2's initial obsturent is underlyingly /h/ (which would be [b] on the surface) in nonce surname judgments (see the experimental results in Section 4.3.4.1). In the MaxEnt model proposed here, a similar interaction effect can be achieved by assigning a relatively low constraint weight to *#b; violations of OCP(voice) and *#b will count less than violations of OCP(voice) and other *#D, because of lower penalty scores they assign to candidates.

The constraints in (92) are all motivated independently of rendaku in surnames. As mentioned above, OCP(voice)-stem drives the effects of normal Lyman's Law observed in rendaku in regular compounds. Moreover, it captures the general scarcity of native stems containing multiple voiced obstruents in the lexicon (see Morita 1977; Yamaguchi 1988; Ito and Mester 1986; also see Section 2.2.2). *#D has also been proposed to be part of Japanese phonology. Generally speaking, native

[6]Cumulative constraint interactions are also known as "ganging": multiple violations of constraints with lower weights may gang up and overturn the effect of a single (or fewer) violation of a constraint with a higher weight. Constraint ganging is one of the characteristics which differentiate Harmonic Grammar models, including MaxEnt HG, from other models. For discussion, see Hayes and Wilson 2008 and Pater 2009 and references therein.

stems do not begin with a voiced obstruent (Martin 1987:30), which can be traced back to the patterns in Old Japanese (Hashimoto 1938/1950; Unger 1977).[7] Thus, all these constraints play a general role in the language.

I also propose the constraints in (93) which formalize the effects of avoidance of stem-internal hormorganic voiceless obstruent sequences and sequences of identical moras.

(93) Constraints on Identity Avoidance

i. *s-s:
Assign one violation mark for every pair of [s] and [s] occurring in adjacent syllables within a stem

ii. *t-t:
Assign one violation mark for every pair of [t] and [t] occurring in adjacent syllables within a stem

iii. *k-k:
Assign one violation mark for every pair of [k] and [k] occurring in adjacent syllables within a stem

iv. *s-t:
Assign one violation mark for every pair of [s] and [t] occurring in the order of [s] followed by [t] in adjacent syllables within a stem

v. *t-s:
Assign one violation mark for every pair of [t] and [s] occurring in the order of [t] followed by [s] in adjacent syllables within a stem

vi. *MORAICIDENTITY (*MORAID):
Assign one violation mark for every sequence of identical moras within a stem

There is also independent evidence for the presence of this family of constraints in Japanese phonology. It has been shown that native Japanese stems with homorganic voiceless obstruents in adjacent syllables are generally underrepresented in the lexicon (Kawahara et al. 2006). These stem-internal restrictions affect rendaku application in compound surnames due to their representations as single stems. (See Section 3.3.2 for the details.) The constraints *s-s, *t-t, and *k-k ban sequences of identical voiceless obstruents, while *s-t and *t-s ban coronal obstruent sequences

[7] See Westbury and Keating (1986) for a phonetic motivation of the restiction on voiced obstruents in prosodically-initial position in general.

with a manner mismatch. Proposing a constraint for each order of [s] and [t], as in *s-t and *t-s, is to capture the asymmetry between the /s+t/ condition and the /t+s/ condition observed in both the corpus study and the experiment; rendaku is promoted in /s+t/ but not in /t+s/ (see Section 4.3.5).[8] It is thus expected that the constraint *s-t will have a higher weight than *t-s in the model. *MORAICIDENTITY, which penalizes surnames with a sequence of identical moras as in [seka-kawa] 'nonce-river,' is included in the model to reflect the fact that the avoidance of total moraic identity has a stronger effect than avoidance of consonantal identity (see the experimental results in Section 4.3.4.1; also see Kawahara and Sano 2014a, 2016).

How these constraints are violated is illustrated in the tableau in (94).

(94) Violations of constraints on Identity Avoidance

Input	Output	*s-s	*t-t	*k-k	*t-s	*s-t	*MORAID
/nisi-ℜ-sawa/	(a) nisi-sawa	*					
	(b) nisi-zawa						
/matu-ℜ-ta/	(c) matu-ta		*				
	(d) matu-da						
/taki-ℜ-kawa/	(e) taki-kawa			*			
	(f) taki-gawa						
/matu-ℜ-sima/	(g) matu-sima				*		
	(h) matu-sima						
/masu-ℜ-ta/	(i) masu-ta					*	
	(j) masu-da						
/naka-ℜ-kawa/	(k) naka-kawa			*			*
	(l) naka-gawa						

Let us now discuss what constraints would be necessary to capture the rendaku patterns of surnames with a sonorant in E1. The results of the corpus study have revealed that surnames with E1-/n/ have relatively high rendaku rates while those with E1-/r/ have low rates. By contrast, the experimental results have shown that none of the sonorant types in E1 significantly affects rendaku application rates. I have argued that the discrepancy is due to the unnaturalness of the patterns; the

[8] I admit that this is merely a post-hoc justification made based on the observations of the corpus and experimental results. I leave a theoretical analysis of the asymmetry for future research.

effects of E1-sonorants observed in real surnames are language-specific and lack clear phonetic motivation. Japanese speakers do not learn such unnatural phonological patterns, and thus the effects are not replicated in a productivity test of rendaku using nonce surnames. (See Section 4.3.5 for the details.) Since the goal of the study is to construct a grammar model which accounts for the phonological behaviors of present-day Japanese speakers, I propose that the model should not include any constraints on sonorants and voicing (e.g. *nVT, *rVD). I will come back to the issue in the next section and discuss how the model is still able to capture the sonorant-driven rendaku patterns in real surnames without relevant phonological constraints.

Lastly, I propose the constraints in (95) to capture the effects of special moras and the length of E1. Note that these constraints will be important to the modeling of rendaku in real surnames but will not be directly relevant to the modeling of rendaku in nonce surnames since the experimental data do not include surnames with a special mora in E1 nor those with one-mora or three-mora E1 (that is, E1 is always bimoraic).

(95) Constraints for other patterns

i. POST-NASALVOICING (*NT):
Assign one violation mark for every sequence of a nasal consonant followed by a voiceless obstruent

ii. NOVOICEDGEMINATE (*DD):
Assign one violation mark for every voiced geminate obstruent

iii. REALIZE-MORPHEME-1μ-E1:
Assign one violation mark for every linking morpheme [+voice]$_{\mathfrak{R}}$ which is not realized on the surface in a surname with a monomoraic E1

iv. REALIZE-MORPHEME-3μ-E1:
Assign one violation mark for every linking morpheme [+voice]$_{\mathfrak{R}}$ which is not realized on the surface in a surname with a trimoraic E1

The effects of POST-NASALVOICING (*NT) is observed both in Japanese (see e.g. Ito and Mester 1995b) and cross-linguistically (see Pater 1999 and references therein), and the constraint is grounded in phonetics (Hayes and Stivers 1996). NOVOICEDGEMINATE (*DD) has also been claimed to be operative in Japanese phonology (see Ito and Mester 1995b, 1999; Nishimura 2003; Kawahara 2006, 2008, 2011 among others), and is motivated by aerodynamic difficulties (see

Kawahara 2006 and references therein). Recall that moraic nasals in E1's last mora promote rendaku voicing of the following obstruent while gemination inhibits the voicing alternation (see Section 4.2.4.3). *NT and *DD are thus expected to have relatively high weights, behaving as a rendaku trigger and as a rendaku blocker respectively. Vowels in E1's last mora do not particularly affect rendaku application. Thus, I do not include in the model any constraints on the relation between a vowel and the following obstruent's voicing.

The corpus study has also revealed the effects of the length of the first element; one-mora E1 and three-mora E1 generally promote rendaku. As discussed above, this is not unique to surnames, and I do not have a proper account of the patterns. Here, I propose two ad-hoc constraints which simply translate the length effects; REALIZE-M-1μ-E1 and REALIZE-M-3μ-E1 require that rendaku voicing be realized specifically when the first element is monomoraic and bimoraic respectively. Further investigation is needed to settle the issue of how these effects can be formally derived.

In the next subsection, I will extend some of the general constraints proposed above and make them specific to E2 morphemes or to compound surnames in order to capture the lexical factors in rendaku in surnames.

5.3.3 Lexically-specific constraints

I have proposed general constraints to be included in the MaxEnt model to account for the rendaku patterns of compound surnames involving free variation. As discussed in Section 5.2, however, rendaku in surnames also show lexical variation. In order to capture their behaviors constrained by lexicalization, I will further propose lexically-specific constraints along the lines of Rosen (2016) and Moore-Cantwell and Pater (2016).

First, I propose sub-versions of REALIZE-MORPHEME and IDENT(voice) which are specific to every E2-morpheme. That is, the two constraints are indexed with each of the morphemes appearing as the second element in the data. There are 122 unique E2 morphemes in the real name data. I thus propose to include 122 REALIZE-MORPHEME constraints and 122 IDENT(voice)

constraints, with each of them being indexed with one E2 morpheme.[9] Some examples of these E2-specific constraints are given in (96) below.

(96) E2-specific constraints on rendaku

i. REALIZE-MORPHEME-E2-/ta/:
Assign one violation mark for every linking morpheme [+voice]$_{\text{R}}$ which is not realized on the surface in the compound surname with /ta/ as E2

ii. REALIZE-MORPHEME-E2-/kawa/:
Assign one violation mark for every linking morpheme [+voice]$_{\text{R}}$ which is not realized on the surface in the compound surname with /kawa/ as E2

iii. IDENT(voice)-E2-/ta/:
Assign one violation mark for every output segment that differs in the feature [voice] from its input correspondent in the morpheme /ta/

iv. IDENT(voice)-E2-/kawa/:
Assign one violation mark for every output segment that differs in the feature [voice] from its input correspondent in the morpheme /kawa/

These E2-specific constraints are expected to capture the irregularities in rendaku application caused by the lexical propensities of E2 morphemes. As mentioned above, morphemes such as /kuti/ 'entrance, gate' undergo voicing in most of the surnames they occur in as the second element. This may even overturn the effects of Strong Lyman's Law, which would otherwise be a strong rendaku-blocking factor, as in/mi̲zo-kuti/ [mi̲zo-**g**uti] 'trench-entrance.' To model such lexically-conditioned patterns, we can give high constraint weights to the versions of REALIZE-MORPHEME indexed with those E2 morphemes. The constraints will then have strong effects in the grammar, causing rendaku voicing in most surnames containing those E2 "rendaku lovers" (Rosen 2001). On the other hand, morphemes such as /kura/ 'warehouse' resist rendaku in most of the surnames they occur in. Constraints like REALIZE-MORPHEME-E2-/kura/ are thus expected to have high weights, inhibiting rendaku application in surnames with those "rendaku haters" as E2. Put simply, E2-specific constraints capture the baseline of every E2 morpheme with

[9]The count of E2 is based on kanji. There are some morphemes that are phonologically and semantically the same but are written with different characters, as in /sima/ 'island' 島 or 嶋. Since the household numbers of surnames in the database are kanji based, these E2 morphemes were treated as different morphemes.

respect to rendaku application.[10]

The idea of incorporating the idiosyncratic properties of every E2 morpheme regarding rendaku voicing into the grammar is not new. Rosen (2016) gives an analysis of rendaku in regular compounds under Gradient Symbolic Computation (Smolensky and Goldrick 2015), a version of Harmonic Grammar (Legendre et al. 1990; see above). This model not only includes weighted constraints but also allows phonological features to have continuous activation levels, which determine the possibility of their surface realization. Rosen proposes that, in compound formation, the elements of a compound (i.e. E1 and E2) each activate the [+voice] feature of an affix that links the two elements (the equivalent of [+voice]$_{\text{R}}$ in this study) to a certain degree. Every morpheme has some idiosyncratic voicing activation value, and rendaku voicing is realized when the effects of E1 and E2 accumulate to cause the affix [+voice] to exceed some threshold activation level. In other words, Rosen's (2016) analysis attributes rendaku application in regular compounds almost entirely to the lexical properties of E1 and E2 and their interactions (which are incorporated in the grammar system itself). The MaxEnt model with E2-specific constraints proposed above is similar to Rosen's Gradient Symbolic Computation grammar except that the model here expresses element-specific properties through constraints and their weights, and does not take E1's idiosyncraic behaviors into account.[11] (See Section 5.5 for other differences between Rosen's model and the MaxEnt model argued for in this dissertation.) As is claimed by Rosen (2016), a full analysis of the phenomenon of rendaku in regular compounds must somehow take the idiosyncrasies of compound elements into account. I will show that the same is true for rendaku in surnames.

Second, I also propose that REALIZE-MORPHEME and IDENT(voice) be indexed with every single surname. See (97) below for examples of the surname-specific constraints.

[10]This is parallel to the random intercepts in the mixed-effects logistic regression models shown in Chapter 4.

[11]E1's lexical properties are not incorporated into the model here since it already includes consonants in E1 as important factors. Having both general phonological constraints and E1-specific constraints would have redundancies and would also increase the possibility of overfitting to training data (see Section 5.3.4).

(97) Surname-specific constraints on rendaku

i. REALIZE-MORPHEME-/naka-ℜ-ta/:
Assign one violation mark for every linking morpheme which is not realized on the surface in the compound surname /naka-ℜ-ta/

ii. REALIZE-MORPHEME-/naka-ℜ-kawa/:
Assign one violation mark for every linking morpheme which is not realized on the surface in the compound surname /naka-ℜ-kawa/

iii. IDENT(voice)-/naka-ℜ-ta/:
Assign one violation mark for every output segment that differs in the feature [voice] from its input correspondent in the surname /naka-ℜ-ta/

iv. IDENT(voice)-/naka-ℜ-kawa/:
Assign one violation mark for every output segment that differs in the feature [voice] from its input correspondent in the surname /naka-ℜ-kawa/

As discussed above, a standard MaxEnt Harmonic Grammar model with general constraints would be appropriate for modeling phenomena involving token (free) variation but not for modeling those showing so-called lexical variation or lexically-conditioned exceptional patterns. Moore-Cantwell and Pater (2016) tackle the problem by combining MaxEnt HG with lexically-specific constraints. In their analysis of stress assignment in English (see above), they propose that there are two general constraints, the one triggering penultimate stress (e.g. *banána*) and the other triggering antepenultimate stress (e.g. *Cánada*), and that every lexical item is associated with lexically-specific versions of those. They show that a MaxEnt model with general and lexically indexed constraints can account for not only the pronunciations of nonce words with token variation observed in an experiment but also the fixed pronunciations of individual items in the lexicon.[12]

Along the lines of Moore-Cantwell and Pater (2016), I propose that each compound surname is associated with one rendaku-triggering constraint (REALIZE-MORPHEME) and one rendaku-

[12]One may think that indexing constraints to classes of words based on their stress patterns (i.e. the penulimate group and the antepenultimate group) would suffice in the case of English stress. However, class based indexation would not be able to capture the fact that the number of items in each class affects the gradient productivity in nonce word experiments. For example, speakers of a language with a stress system similar to English but with more words with antepenultimate stress and fewer words with penultimate stress (as "exceptional patterns") would use antepenultimate stress more often in the variable pronunciations of nonce words, showing frequency match to the lexicon. Moore-Cantwell and Pater 2016 claim that constraints indexed with each lexical item is necessary in order to capture this kind of "gradient exceptionality." See Moore-Cantwell and Pater 2016 for more details.

blocking constraint targetting (IDENT(voice)). The job of these constraints is to regulate the idiosyncrasy of each surname. As we have seen earlier, surnames with a /k...k/ sequence generally undergo rendaku, but there are some individual items which are commonly pronounced without voicing. Examples include [seki-kawa] 'barrier-river' and [yoko-kawa] 'side-river.' The behaviors of these surnames cannot be explained in terms of phonological constraints (/k...k/ promotes rendaku) nor in terms of E2-specific constraints (/kawa/ is a rendaku waverer). Surname-specific constraints like IDENT(voice)-/seki-ℜ-kawa/ and IDENT(voice)-/yoko-ℜ-kawa/ will then come into play (having a relatively high weight) and block rendaku in these particular surnames. There are 1064 unique surnames in the real name data. The proposal is thus to include 1064 REALIZE-MORPHEME constraints and 1064 IDENT(voice) constraints, with each of them being indexed with each surname. (See Section 5.5 for discussion on whether the constraint set is reasonable in size.)[13]

Crucially, non-existing surnames are not affected by these constraints which are indexed to existing surnames. In the case of nonce surname, rendaku is determined through interactions of general and E2-specific constraints, which yield token variation. Thus, what is important is to let the model capture the fixed patterns of real surnames using lexically-specific constraints but still have it learn the overall rendaku patterns based on phonological factors and the lexical propensities of E2-morphemes so that it will make the right predictions about novel surnames. In the next section, I propose implementing learning biases on constraints using a Gaussian prior in order to prevent the model from overweighting surname-specific constraints.

[13]It is important to include lexically-specific versions of both REALIZE-MORPHEME and IDENT(voice) in the model. Take, for example, surnames with a sonorant in E1. There are no general phonological constraints that target them. We want the model to assign similar weights to general REALIZE-MORPHEME and IDENT(voice) to capture the rendaku patterns of nonce surnames with a sonorant in E1, since they undergo rendaku almost at a chance level (see Section 4.3.4.2). However, real surnames with a sonorant in E1 show a wide range of rendaku rates; some always undergo rendaku (e.g. [kuro-**d**a] 'black-paddy') while others never do (e.g. [kuri-**t**a] 'chestnut-paddy'). Thus, to capture the behaviors of rendaku lovers, lexically-specific REALIZE-MORPHEME must have high weights, and to capture the behaviors of rendaku haters, lexically-specific IDENT(voice) must have high weights.

5.3.4 Implementing learning biases

As stated above, the goal of the learning algorithm of MaxEnt HG is to find the constraint weights that maximize the probability of the observed data. This enable us to further examine how a probabilistic model that best fits the existing data can make predictions about the phonological patterns of novel words. However, the problem of such a maximum likelihood learning model is potential overfitting to the training data. If the learner is given a finite sample of data and simply asked to find the best fit, it may i) overestimate the probability of the items in the provided sample, and ii) underestimate the probability of other real items which did not happen to occur in those data or the probability of novel items which could possibly occur (and which are in theory infinite). The problem of overfitting can be especially significant for a model like the one proposed here which contains a large set of lexically-specific constraints. Given constraints associated uniquely with every item, a model can easily achieve the maximum likelihood of the observed data by assigning very high weights to those lexically-specific constraints. Such an outcome is not ideal, since we would like the model to also make use of phonological constraints and E2-specific constraints in order to predict the rendaku patterns in non-existing surnames.

The method commonly used to avoid such overfitting in a MaxEnt model is to introduce a regularizing bias term, or a "prior," into the learning function. In the MaxEnt model used in this dissertation, I use a Gaussian prior; the prior for each constraint weight w_j is a Gaussian distribution with mean μ_j and a standard deviation σ_j, as shown in (98).

(98) The prior term

$$\sum_{j=1}^{m} \frac{(w_j - \mu_j)^2}{2\sigma_j^2}$$

The prior term is then subtracted from the log probability of the data in (88), as shown in (99).

(99) The MaxEnt learing function with a Gaussian prior

$$\sum_{i=1}^{n} logP(y_i|x_i) - \sum_{j=1}^{m} \frac{(w_j - \mu_j)^2}{2\sigma_j^2}$$

With the inclusion of a Gaussian prior, a MaxEnt model now aims to find the constraint weights which maximize this objective function. More intuitively, the prior in the function can be understood as a bias or a penalty imposed on constraint weight learning. μ in the formula functions as a parameter for the preferred weight of each constraint. As can be seen, the μ value is subtracted from the learned weight w, and the resulting value is squared. This means that the farther the constraint weight is from its expected value (i.e. μ), the greater penalty is imposed on the learning. σ, on the other hand, functions as a parameter for how much the constraint can deviate from its ideal weight. The square of the difference between the weight and its preferred value is divided by two times the square of σ. This means that the lower the σ^2 value is, the greater the penalty is, with generally more data being required for the weight to move away from μ.

In the implementation of my model for rendaku in surnames, I set the μ value of every constraint to be zero. This means that the ideal weight of every constraint is zero and any weight higher than that will receive some penalty. In practical terms, if the priors are set uniformly low, the model generally assigns low weights to all the constraints. For example, if multiple constraints turn out to be able to explain the same aspect of the learning data, the model tries to distribute weights to those relevant constraints, instead of assigning a very high weight to one or a few of them merely to achieve the maximum likelihood of the observed data in a literal sense (as defined in (88) above). Thus, the prior generally helps avoid potential overfitting because of the large number of lexically-specific constraints (see the discussion above).

It is also important to note that the prior does not actually operate uniformly on every constraint, in the sense that some are affected more than others. For example, a constraint indexed to a single lexical item (e.g. IDENT(voice)-/saka-ℜ-ta/) is only capable of explaining the sound pattern of that item (*[saka-**d**a]), while a general constraint (e.g. IDENT(voice)) accounts for the patterns of many items that incur its violation (*[saka-**d**a], *[yama-**g**awa], *[taka-**z**ima], etc.). This means that,

between general constraints and lexically-specific constraints, the former simply have more tokens in the learning data allowing them to achieve weights which are far away from their preferred weight. General constraints are thus expected to receive relatively high weights. In other words, setting low μ values on every constraint translates to a kind of generality bias, which has long been claimed to be essential in the formalization of phonological patterns (see e.g. Chomsky and Halle 1968 for an argument for formal simplicity in their evaluation metrics). This fact also helps the model learn the rendaku patterns based on general phonological factors, which are important in predicting rendaku application in nonce surnames, rather than on lexical factors.

I opt to set the σ values of all constraints at thirty. Relatively speaking, this number is neither high nor extremely low. Again, σ allows for adjusting the effects of the priors, which are important for our purposes since they make differences in the learning of general constraints versus that of lexically-specific constraints. Put another way, changing σ values is equal to changing the balance between the role of general phonological factors and the role of lexicalization in the grammar, which will prove crucial in the modeling of the rendaku patterns of surnames. I will come back to this discussion in Section 5.4.3, in which I compare models with different σ values.

For the implementation of the model, I used the MaxEnt Grammar Tool (Wilson et al. 2006).

5.3.5 Summary

I have proposed a grammar model to account for rendaku application in Japanese surnames. Maximum Entropy Harmonic Grammar is used as the theoretical framework, which can learn the token-based variability in the rendaku data and make predictions about the voicing alternation in nonce surnames. I have proposed constraints which reflect the effects of phonological factors, such as Strong Lyman's Law and Identity Avoidance, following the results of the corpus study and the experiment. I have also proposed lexically-specific constraints in order to capture the lexical variation in the data: constraints which are indexed to each E2 morpheme and those which are indexed to each compound surname. I have argued that, with learning biases, the model will make the right predictions about the phonologically-driven rendaku application patterns as well as the lexically-

conditioned patterns.

5.4 Testing the model

In this section, I test the predictions of the proposed MaxEnt Grammar through learning simulations. First, I will train the model with the data of rendaku in existing surnames obtained from the corpus study. The model's fit to the training data will then be assessed. Second, I will test the model's predictions against the rendaku patterns of non-existing surnames observed in the experiment. It will be shown that the model correctly accounts for both the lexicalized rendaku patterns in the corpus and the productive rendaku patterns in the experiment. I will also compare the proposed MaxEnt grammar with alternative models and discuss the necessity of lexically-specific constraints and learning biases.

5.4.1 Predicting rendaku in real surnames

For the training data, the rendaku patterns of 1064 existing surnames obtained in the corpus study were used. A MaxEnt Grammar which consists of the general and lexically-specific constraints described above was given the data of real surnames with rendaku rates. In the learning, the model considered the rendaku form and the non-rendaku form as possible outputs for each input surname. For example, for the surname /naka-ta/, [naka-ta] without rendaku and [naka-da] with rendaku were the two candidate output forms. As discussed above, the μ of every constraint was set to be zero and the σ was set to be thirty.

I present in (100) the weights of constraints that the model has learned from the training data. Due to limitations of space, only the forty-five constraints with the highest weights are shown. E2-specific constraints are underlined and general constraints are both underlined and boldfaced. REALIZE-MORPHEME and IDENT(voice) are abbreviated as "RM" and "IdV" respectively. $\mathfrak{R}$ in surname-specific constraints is omitted for simplicity.

(100) The top 45 constraints in the model

Constraint	Weight	Constraint	Weight	Constraint	Weight
IdV-/se-ta/	3.83	IdV-E2-/take/	2.76	RM-/soe-sima/	2.53
RM-/matu-hara/	3.81	RM-/tama-kawa/	2.74	RM-/sugi-saki/	2.52
IdV-/se-to/	3.76	IdV-E2-/se/	2.73	RM-/oka-sima/	2.52
IdV-/mina-kuti/	3.68	RM-/kita-kawa/	2.68	RM-/miya-sima/	2.51
RM-E2-/hayasi/	3.49	RM-/ima-kawa/	2.68	IdV-/ara-kaki/	2.50
RM-/do-hasi/	3.40	IdV-E2-/ki/	2.68	RM-/kita-sima/	2.49
IdV-/su-saki/	3.40	RM-E2-/kuti/	2.64	IdV-/mi-to/	2.49
RM-/matu-saka/	3.36	RM-/isi-hasi/	2.64	IdV-/ki-kuti/	2.48
IdV-E2-/kura/	3.15	**OCP(voi)**	**2.63**	IdV-/atu-ta/	2.48
IdV-E2-/hara/	2.91	IdV-/to-tuka/	2.63	IdV-E2-/sita/	2.48
IdV-/mu-ta/	2.89	IdV-/a-sato/	2.60	RM-/kasiwa-ki/	2.46
RM-/yana-kawa/	2.77	RM-/kinu-kawa/	2.58	**NoVoiGem**	**2.45**
RM-/sio-hara/	2.77	RM-/miya-sima/	2.54	RM-E2-/he/	2.45
RM-/sina-kawa/	2.77	RM-/ii-sima/	2.53	RM-/sasa-kawa/	2.45
IdV-/kon-ta/	2.76	RM-/huna-hasi/	2.53	IdV-E2-/kisi/	2.44

As can be seen, most of the highly weighted constraints are surname-specific ones. This means that the model has learned the lexically-conditioned rendaku patterns of existing surnames to some degree. Note, however, that the model has also assigned relatively high weights to some of the general constraints, such as OCP(voice). That is, it has learned not only lexicalized patterns but also general phonological patterns. As will be shown below, this proves crucial in the modeling of rendaku in non-existing surnames. (See (101) below for the weights of other general constraints.)

Let us now compare the rendaku patterns of real surnames in the corpus and those predicted by the MaxEnt model. Figure 5.1 plots the average rendaku rates of surnames with an obstruent (/s/, /t/, /k/, or a voiced obstruent represented by captial /D/) in the first element altogether; the graph on top represents the model's predictions while the graph at the bottom shows the results of the corpus study. (As in the graphs shown in Chapter 4, the error bars in the graphs represent ±1 standard errors.) The two graphs look almost identical to each other. This confirms that the model has successfully learned the lexically-conditioned rendaku patterns of existing surnames.

Let us see the model's predictions on the rendaku patterns of surnames with a sonorant or

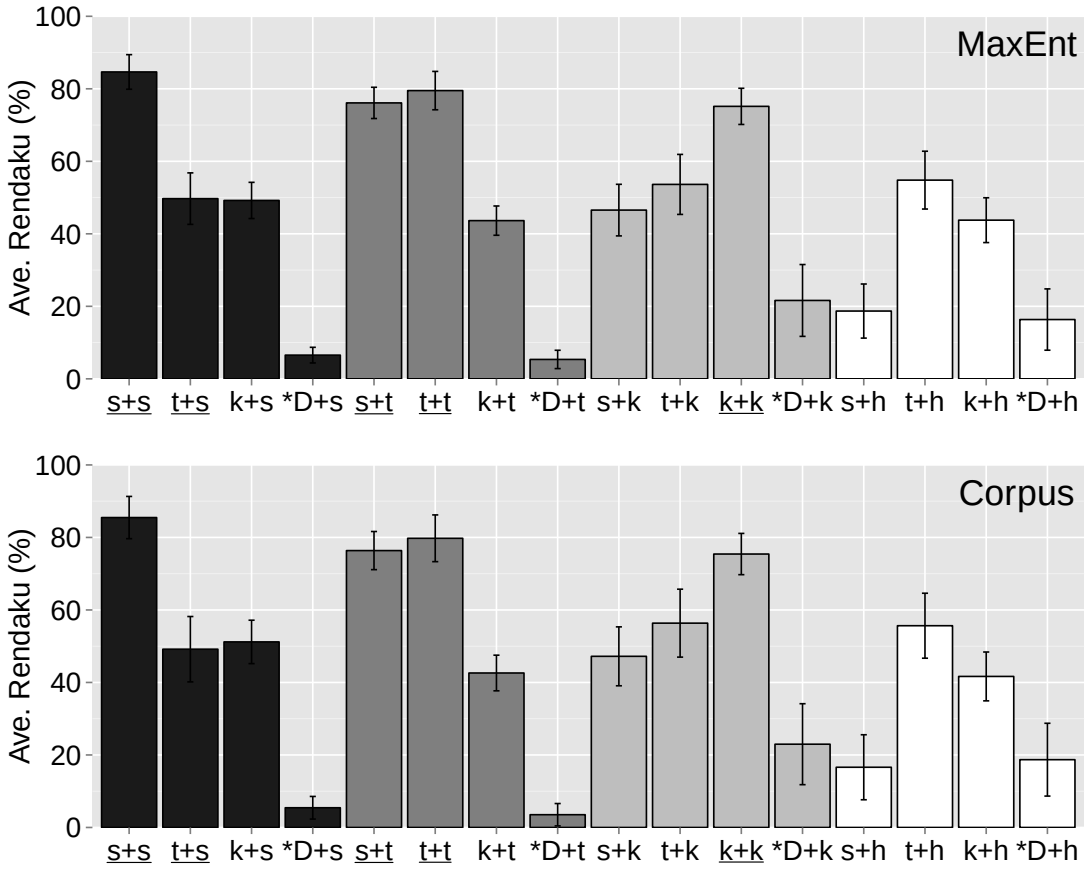


Figure 5.1: Average rendaku rates of surnames with E1-obstruents: MaxEnt vs. Corpus

a special mora in E1. Figure 5.2 compares the MaxEnt predictions (the graph on top) and the corpus results (the bottom graph). The results are shown by consonants in E1, rather than E1-E2 consonant combinations, as in the presentation of the corpus results in Section (Section 4.3.5) due to the general scarcity of the data. "N" represents moraic nasals and "Gem" refers to gemination. Here again, the MaxEnt model makes very accurate predictions about the rendaku patterns of existing surnames.

The success in modeling these names is particularly meaningful. Recall that I did not include any particular phonological constraints which would explain (non-)application of rendaku affected by sonorants in E1 (e.g. *NVT, *rVD), arguing that those patterns are not rooted in real phonological factors (at least synchronically). The model still captures the data very well, suggesting that lexically-specific constraints play a big role in the learning. Given constraints specific to every

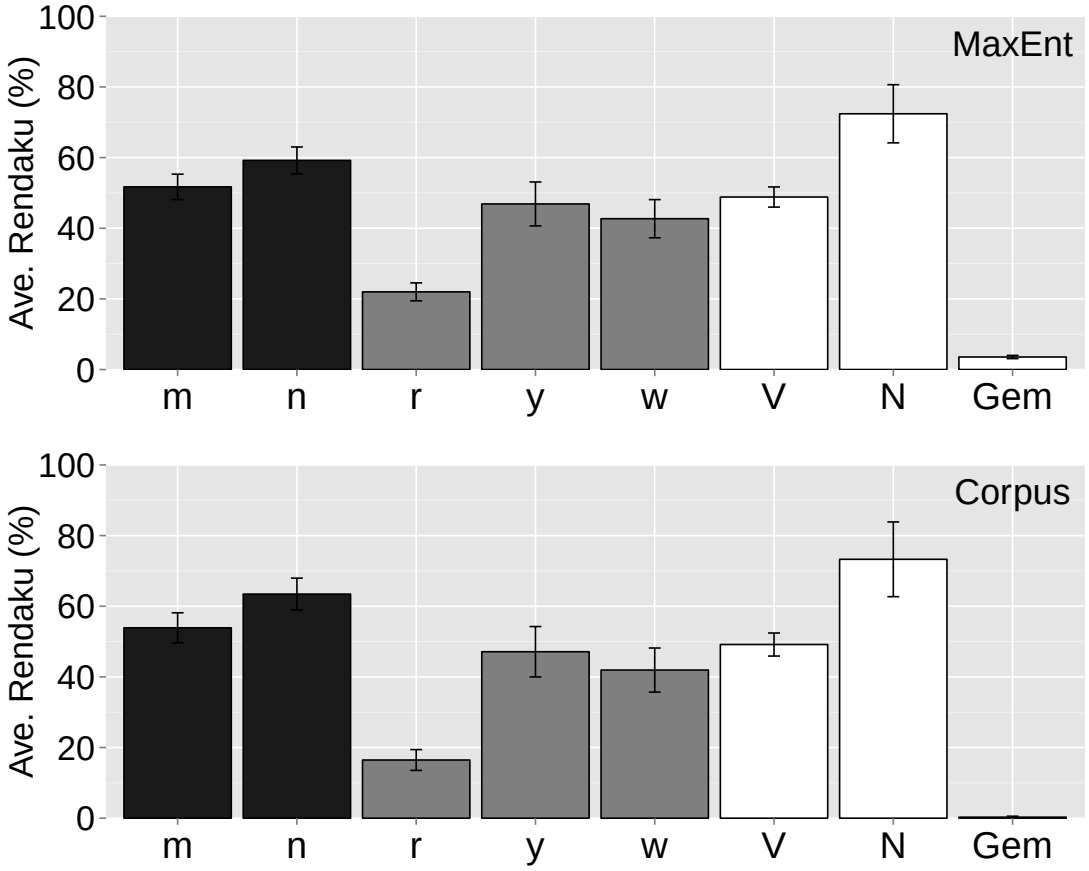


Figure 5.2: Average rendaku rates of surnames with sonorants and special moras in E1: MaxEnt vs. Corpus

surname, a learner does not need to resort to unmotivated constraints of the kind discussed above (which are most likely not to be innate) in order to account for some patterns which are unnatural but do exist in the lexicon. I argue that this actually reflects the learning of rendaku in surnames by Japanese speakers, and that is why these patterns are not productive, as we have seen in the experiment (Section 4.3.5).

Lastly, let us make a more detailed comparison by checking the model's predictions against each data point in the corpus. Figure 5.3 shows a scatter plot of the rendaku rates which MaxEnt predicts by the rendaku rates of all the 1064 surnames obtained from the corpus study. The predictions of the model are very accurate, showing a considerably high coefficient of determination (r^2=0.9822) in relation to the actual rendaku rates. This again corroborates the argument that the

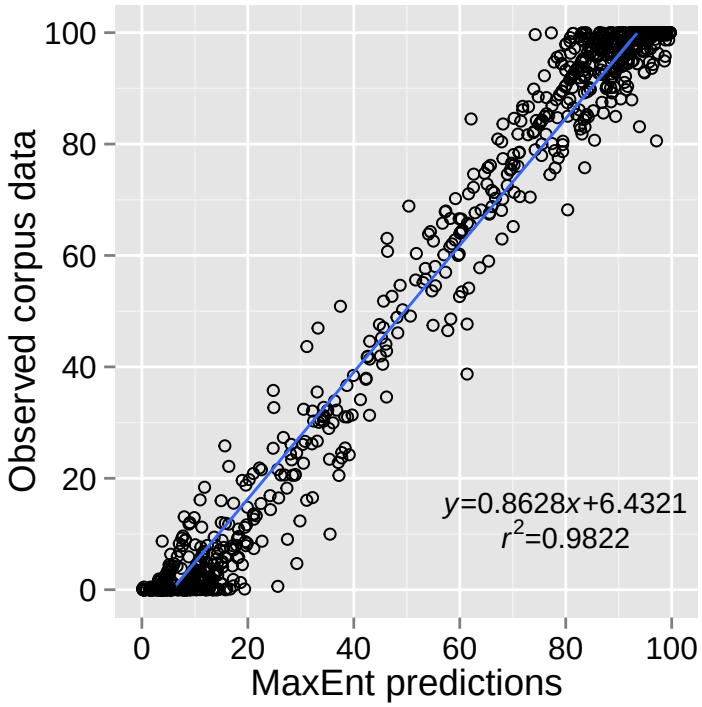


Figure 5.3: Rendaku rates of 1064 existing surnames: MaxEnt vs. Corpus

model can learn the sound patterns of existing surnames. These results may not be too surprising given that I included a large number of lexically-specific constraints in the model. However, explaining the highly lexicalized voicing alternation patterns in terms of a grammar is meaningful, since very few attempts of this kind have ever been made in the long history of the study of the phenomenon (cf. Rosen 2016). As will be discussed later, proposing a grammar with lexically-specific constraints has theoretical advantage over putting the burden of explanation on the lexicon (i.e. simply stating that the patterns are lexicalized). In the next subsection, I will further show that the model is also capable of learning general phonological patterns which are crucial in capturing rendaku application in nonce surnames.

5.4.2 Predicting rendaku in nonce surnames

We have seen that the proposed MaxEnt model is successful at learning and reproducing the rendaku patterns of existing surnames. How does the model perform in predicting rendaku in

non-existing surnames? It should be noted that the training data do not include non-existing surnames and that the model must predict their sound patterns on the basis of the constraint weights it has learned from the data of existing surnames, rather than adjusting weights to fit to the new data. Recall also that constraints indexed to real surnames, which have gained high weights in the learning, do not affect nonce surnames. When those constraints are inactive, general constraints and E2-specific constraints come into play and determine rendaku application in novel items.

With this in mind, let us see whether or how the model has assigned weights to constraints other than surname-specific ones. (101) gives the weights of all seventeen of the general constraints included in the model.

(101) Weights of general constraints

Constraint	Weight	Constraint	Weight
OCP(voice)	2.627	*#z	1.380
NoVoiGeminate	2.451	*#d	1.265
MoraId	2.319	*t-t	1.055
*s-s	2.256	Realize-M	0.157
*NT	1.980	*#b	0.005
*s-t	1.771	Ident(voice)	0.000
Realize-M-3μ-E1	1.689	*t-s	0.000
Realize-M-1μ-E1	1.651	*#g	0.000
*k-k	1.556		

We see that the model did assign some weights to most of the general constraints. These constraints are thus expected to affect rendaku application in nonce surnames. For example, the high-weighting of OCP(voice) will result in so-called Strong Lyman's Law effects. Constraints such as *s-s, *k-k and *s-t will also produce Identity Avoidance effects.

The other type of constraints which are important in the modeling of rendaku in nonce surnames is E2-specific constraints. Since the non-existing surnames used as the stimuli in the experiment in this study are composed of a nonce morpheme as the first element and a real morpheme as the second element, E2-specific constraints are still relevant to the rendaku patterns of those surnames. As discussed above, those constraints specific to E2-morphemes are expected to capture the base-

line of each E2 for undergoing voicing. The weights of all the nineteen E2-specific constraints are given in (102). Those which received the weight of zero are omitted from the table.[14]

(102) Weights of E2-specific constraints

Constraint	Weight	Constraint	Weight
REALIZE-M-E2-/hayasi/	3.487	REALIZE-M-E2-/sawa/	1.569
IDENT(voice)-E2-/kura/	3.149	IDENT(voice)-E2-/kawa/	1.451
IDENT(voice)-E2-/hara/	2.905	IDENT(voice)-E2-/hasi/	1.353
IDENT(voice)-E2-/se/	2.728	REALIZE-M-E2-/saki/	1.181
IDENT(voice)-E2-/ki/	2.679	REALIZE-M-E2-/ta/	0.922
REALIZE-M-E2-/kuti/	2.644	IDENT(voice)-E2-/hata/	0.671
IDENT(voice)-E2-/sita/	2.480	REALIZE-M-E2-/sima/	0.290
IDENT(voice)-E2-/tani/	2.098	REALIZE-M-E2-/tuka/	0.172
IDENT(voice)-E2-/kosi/	2.047	IDENT(voice)-E2-/kami/	0.018
IDENT(voice)-E2-/saka/	1.637		

E2-morphemes with relatively high weights of REALIZE-M associated with them can be considered to be "rendaku lovers" in Rosen's (2001) terms; the effects of highly-weighted E2-specific REALIZE-M facilitate voicing, setting the baseline for rendaku in names with those E2-morphemes higher. On the other hand, those with relatively high weights of IDENT(voice) correspond to "rendaku haters"; highly-weighted E2-specific IDENT(voice) constraints make the voicing alternation generally less likely in names with those E2-morphemes.

Let us see how the model predicts rendaku application in nonce surnames through interactions of the constraints in (101) and (102). The graphs in Figure 5.4 show the average rendaku rates of surnames with an obstruent in E1 observed in the experiment (the bottom figure) and the rates predicted by the model for the same surnames (the top figure).

Although the predictions are not as accurate as the ones for the training data, we can see that the model still captures the general patterns of rendaku in nonce surnames. It predicts that voiced obstruents in the first element (represented by "D") generally inhibit the voicing alternation due

[14] REALIZE-M and IDENT(voice) are opposing constraints. Since E2's propensities for rendaku are either in the direction of undergoing voicing or in the direction of resisting it, only one of the two constraints associated with each E2 morpheme received some weight and the other received zero.

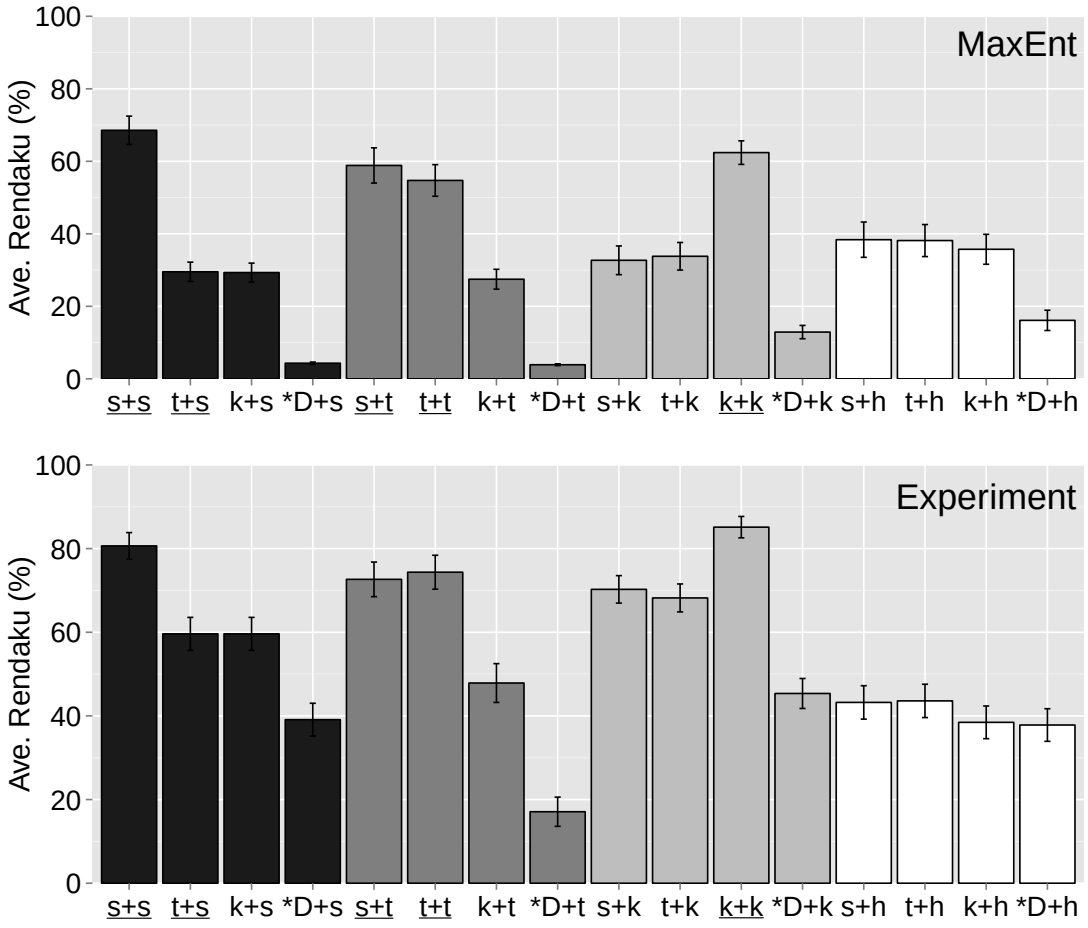


Figure 5.4: Rendaku rates of surnames with E1-obstruents: MaxEnt vs. Experiment

to the high weight of OCP(voice). It also correctly predicts that Identity Avoidance generally promotes rendaku voicing, with constraints such as *s-s, *k-k, *s-t and *t-t being operative. Overall, the rates predicted by the model are lower than those of the experiment. This, however, should not be taken particularly as the model's failure to learn the data patterns. Recall that participants in the experiment generally applied rendaku more often than what was expected from the patterns of existing surnames, possibly due to artifacts of the experimental design (see Section 4.3.5).

Figure 5.5 gives graphs representing the rendaku rates of nonce surnames with a sonorant in E1 predicted by the MaxEnt model (top) and those observed in the experiment (bottom). (The experiment did not include surnames with a special mora in E1, so only surnames with nasals and approximants in E1 are shown.) Since the model does not have general constraints which

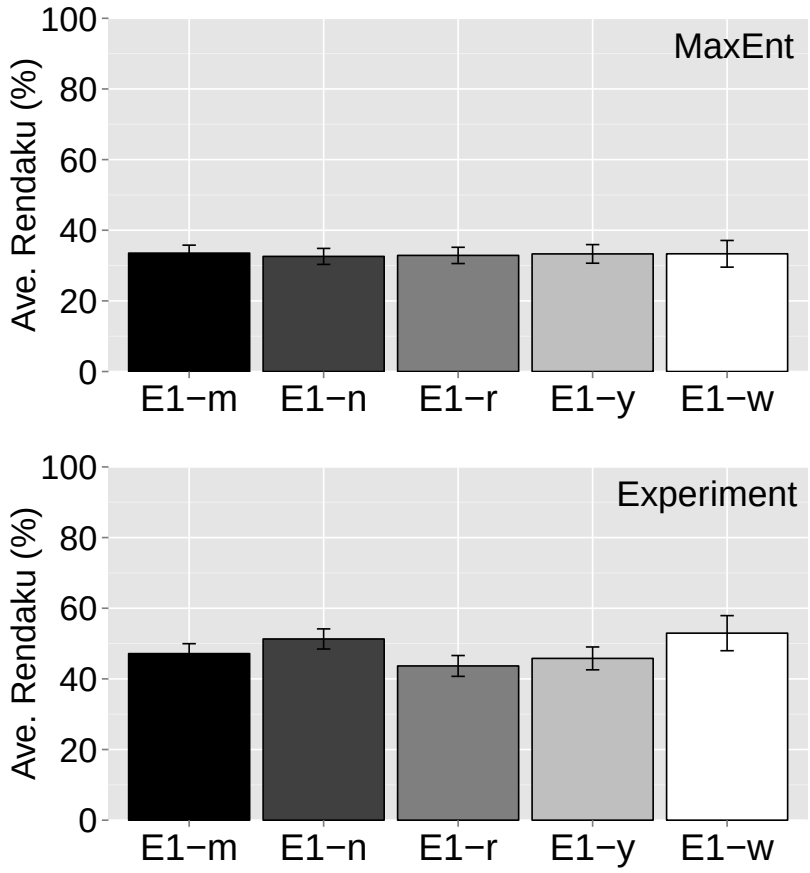


Figure 5.5: Rendaku rates of surnames with E1-sonorants: MaxEnt vs. Experiment

specifically affect rendaku application with a sonorant in E1, rendaku rates are similar among all the conditions. Again, the non-inclusion of such constraints on the relation between sonorants and voicing (e.g. *nVT, *rVD) is meant to reflect the fact that the rendaku patterns conditioned by E1-sonorants are unnatural patterns possibly developed through diachrony. This is compatible with the results of the experiment where none of the E1-sonorant effects observed in existing surnames is attested.

Figure 5.6 gives a scatter plot showing the rendaku rates of 110 types of nonce surnames[15] predicted by the MaxEnt model (x-axis) and those observed in the experiment (y-axis). Again, although the model's predictions are not perfect, they capture the general trend of rendaku application in nonce surnames. It can be seen that the model often underpredicts the rate of voicing.

[15]The types are based on the profiles of constraint violations.

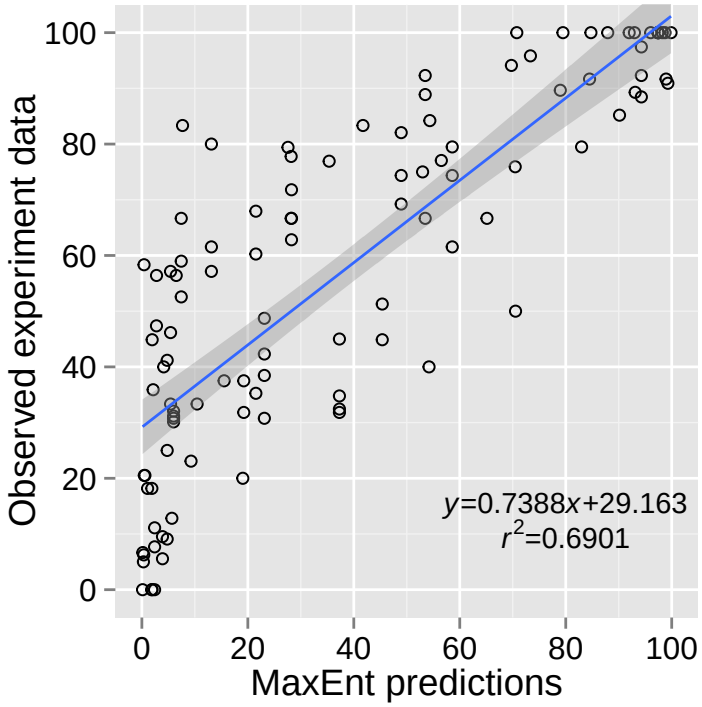


Figure 5.6: Rendaku rates of 110 nonce surname types: MaxEnt vs. Experiment

Again, this is likely due to the relatively high rendaku rates in the experimental results.

Overall, I take the results to be promising. A MaxEnt model composed of a large number of lexically-specific constraints and a small number of general constraints can not only learn the rendaku patterns of existing surnames which are heavily conditioned by lexical factors, but also capture the general patterns of productive rendaku application in non-existing surnames. With biases imposed on the learning of constraint weights, the model assigns some weights to general phonological constraints, which play an important role in the prediction of rendaku in nonce items. In the next subsection, I will discuss the role of the biases in the learning.

5.4.3 The role of lexically-specific constraints and biases

I have claimed that surname-specific constraints are necessary in order to capture the fixed pronunciations of existing surnames (see Sections 5.2 and 5.3.3 for discussion; also see Moore-Cantwell and Pater 2016). For illustration, I compare the performance of the model proposed above and

that of an alternative model. A MaxEnt model was constructed with the same set of constraints but with a σ value of 1 as the prior. As discussed above (Section 5.3.4), lower σ^2 values yield greater penalties for constraint weights which deviate from their expected values (i.e., μ). Also recall that biases generally affect specific constraints more than general constraints, since the former have very few relevant data samples in order for the model to update their weights. Thus, the new model proposed here is a model with stronger learning biases which especially operate against surname-specific constraints.

Let us see how this new model learns the lexically-conditioned rendaku patterns of real surnames. Figure 5.7 shows the rendaku rates of 1064 existing surnames in the corpus by the model's predictions. As can be seen, the model's fit to the training data is considerably worse than that of the original model (Figure 5.3). Although the coefficient of determination is not extremely low (r^2=0.6067), this is partly due to a large number of data points at the extremes. The graph clearly shows that the model fails to capture rendaku application in many of the surnames. The model still does fine in predicting the patterns of nonce surnames (r^2=0.6662), suggesting that it has learned general phonological patterns from the data despite biases on all the constraints. Thus, if learning biases are too strong, they especially affect the surname-specific constraints and render them almost ineffective. As a result, the grammar is incapable of capturing item-based irregularities in the data of real surnames. This provides evidence that surname-specific constraints are crucial in the modeling of rendaku.

However, giving priority to lexically-specific constraints may also cause problems in the modeling of the phenomenon. Suppose there is no bias on constraint weights. The model would then resort to surname-specific constraints more than general constraints during learning. This would make the fit to the real name data even better[16]; yet the model would now be overfitted to the training data and its predictions of rendaku application in non-existing surnames would be less accurate. Therefore, although surname-specific constraints are necessary, it is also important to impose some biases against them so that the model will come to learn general phonological patterns in the given

[16]Indeed, it is possible for a model with surname-specific constraints to achieve an almost perfect fit (e.g. with the coefficient of determination of r^2=0.9999) once the bias level is set very low; e.g. $\mu = 0$, $\sigma = 1000$.

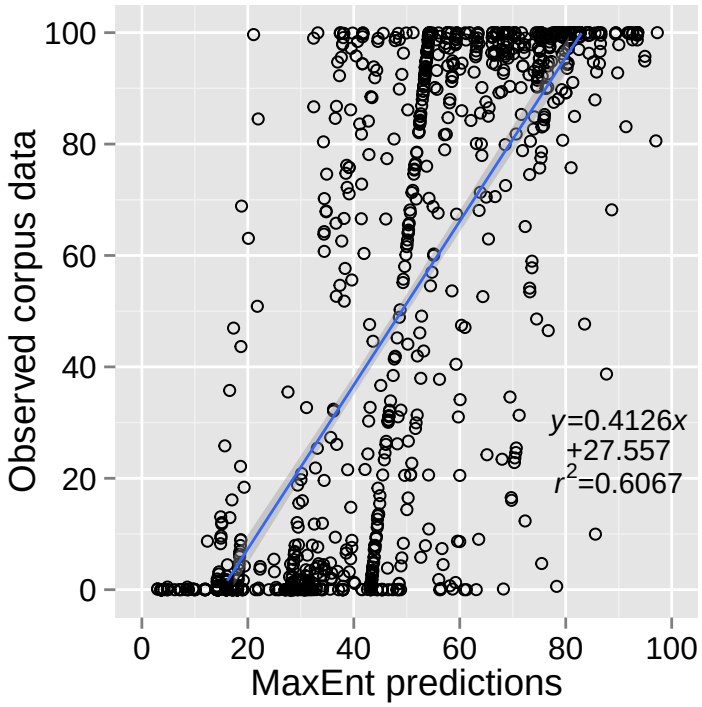


Figure 5.7: Rendaku rates of 1064 existing surnames: MaxEnt with strong learning biases vs. Corpus

data. In other words, there is a trade-off between accurate learning of the lexically-conditioned rendaku patterns of existing surnames on the one hand and accurate predictions of productive rendaku application in non-existing surnames on the other. Thus, adjusting the strength of learning biases (or adjusting σ values in the terms of MaxEnt HG) in the proposed model means finding the right balance between the role of lexical factors and the role of pure phonological factors in the grammar.

5.4.4 Summary of the results

With learning simulations, I have shown that a MaxEnt HG model with general and lexically-specific constraints can not only accurately learn and reproduce the rendaku patterns of existing surnames in the corpus data, but also generate rendaku patterns of non-existing surnames which generally agree with what was observed in the experiment. I have also shown that general con-

straints, lexically-specific constraints and learning biases against them are all needed in order to fully account for the rendaku phenomenon in surnames.

5.5 General discussion

The results of the MaxEnt modeling have several implications for the theory of phonology. First, it demonstrates that a single grammar can account for both highly-lexicalized patterns of some alternation in real words and more phonologically-conditioned productive patterns of the same alternation seen in nonce words. It has been observed in the literature that speakers show "frequency match" based on the distribution of the lexicon of their language in the production or judgment of nonce forms. As discussed above (Section 5.2), a challenge that comes with modeling such a phenomenon is the following. Speakers make generalizations of the lexical distribution based on phonological factors and reflect them in the form of free variation in nonce items; however, the actual pronunciation of each of the existing items is often fixed and should not be produced variably. As Moore-Cantwell and Pater (2016) point out, very few studies have ever attempted to model the lexicalized patterns of real words and the variable patterns of nonce words at a time (cf. Zuraw 2000). Moore-Cantwell and Pater (2016) propose to combine MaxEnt Harmonic Grammar with lexically-specific constraints in order to address the issue. This dissertation offers a large-scale case study testing their proposal. The results are promising in that the grammar model is able to capture the rendaku patterns in both real and nonce surnames.

One may argue that the size of the constraint set in the proposed model is "too large." It does include a relatively large number of constraints; 17 general constraints, 244 E2-specific constraints (of which about one half are really effective with some positive weights) and 2128 surname-specific constraints (of which about one half are actually effective). Note, however, that the MaxEnt model did converge, suggesting that the grammar is feasible in terms of learnability at least according to a learning simulation. It should also be noted that, in order to give a full analysis of rendaku in surnames, the lexical propensities of every E2 morpheme and every compound surname to voice must be captured in some way or another. One alternative to having lexically-specific constraints

would be to incorporate this information in the lexicon itself, with every E2-morpheme and every surname being tagged for how likely they undergo rendaku. This approach, however, does not seem to be better in terms of memory use or learnability. In the absence of better alternatives, the mere fact that the size of its constraint set is large cannot be a strong criticism of the model itself.

Furthermore, incorporating lexical factors into the grammar actually has advantage over encoding rendaku information in the lexicon. A grammar-based explanation of the kind proposed above is a formal analysis and is falsifiable. It can be tested against the patterns of existing words and also the patterns of non-existing words (obtained from experimental studies). By contrast, a simple statement that the voicing status of every surname is stated in the lexicon does not address how the existing patterns would be learned, nor does it provide any explanation as to why rendaku occurs at all in nonce surnames. As I have shown above, an account of rendaku in surnames must take both lexical factors and phonological factors into consideration. The MaxEnt model I have proposed here incorporates the two kinds of factors and even makes explicit how much of a role each of them plays in the form of constraint weights. It would be difficult, if not impossible, for a grammar without lexically-specific constraints to determine what (or how much) is lexicalized and what (or how much) is not lexicalized in a non-arbitrary way.

While introducing lexically-specific constraints, I have mentioned Rosen's (2016) account of rendaku in regular compounds within the framework of Gradient Symbolic Computation, or "GSC" for short (Smolensky and Goldrick 2015), which also incorporates lexical factors into a stochastic grammar (see Section 5.3.3 for the details of the grammar design). One crucial difference between my MaxEnt model and Rosen's GSC model is that the latter explains rendaku application entirely through interactions of the lexical properties of the first element and the second element of a compound, and does not take phonological factors into account. Although the grammar is well defined and does quite a good job in capturing rendaku in existing words, it seems to have some unwanted consequences (some of which are similar to the ones discussed above).

For one thing, it is not entirely clear how Rosen's model would predict rendaku application in novel compounds in which which one of the elements (or possibly both) lacks lexical information about its voicing activation value. For another, it could miss some of the major phonological

generalizations in the language. Although Rosen (2016) excludes E2-morphemes with a voiced obstruent (e.g. /tabi/ 'travel') from analysis, his model in its current design would treat them as having extremely low voicing activation values, which would turn into a rendaku-blocking factor. What they have in common, however, is that they contain a voiced obstruent, and voicing of their initial obstruent would cause a violation of OCP(voice) (e.g. *[**d**abi]), which also has its effects elsewhere in the lexicon (see Section 2.2.2; also see Morita 1977; Yamaguchi 1988; Ito and Mester 1986). The model would thus fail to differentiate the non-application of rendaku due to the lexical propensities of E2-morphemes and that driven by general phonological factors.

To be fair, Rosen's (2016) goal was to model lexically-conditioned rendaku application in existing regular compounds, glossing over the effects of phonological factors (hence the exclusion of E2-morphemes with a voiced obstruent), and he also did not intend to extend the analysis to make predictions about rendaku application in novel compounds. In theory, it is possible to incorporate general phonological constraints such as OCP(voice) in his GSC model as well, in which case it could capture general phonological patterns in existing compounds and also possibly in non-existing compounds. The focus here should instead be on the fact that both lexical and phonological factors are necessary to account for the lexicalized and productive rendaku application patterns when they are both present, as has been argued repeatedly in this chapter. This is especially true in the case of rendaku in surnames. As we have seen, multiple kinds of phonological effects come into play in the voicing alternation in surnames, unlike in regular compounds where such effects are mostly unattested and the lexical properties of words seem to be more important (see Section 4.2.5; also see Irwin 2014a, 2016a). Thus, rendaku in surnames poses an especially interesting problem in that phonological and lexical factors are both crucial and they are mingled together. My MaxEnt model meets the challenge of capturing such patterns with general and lexically-specific constraints and learning biases.

Lastly, the general success in modeling rendaku application is meaningful in the field of Japanese phonology. There have been debates as to whether or not compound voicing should be considered to be a problem in phonological theory (see Kawahara 2015a and references therein for discussion). This is mainly due to the two aspects of the phenomenon which are seemingly

contradictory. On the one hand, its application is irregular and highly conditioned by the lexical properties of given compounds, but on the other hand, it applies productively in newly-coined words just like a normal phonological alternation. In the long history of research on rendaku, the current study offers the very first attempt to tackle the issue by giving a unified formal analysis of both the lexicalized aspect and the productive aspect of the phenomenon. Although the main focus here is on rendaku in compound surnames, it is essentially the same as rendaku in regular compounds in that lexical and productive patterns coexist. In fact, modeling rendaku in surnames is even more challenging in that phonological factors play a big role. The results presented here thus send a strong message as an answer to the long standing problem in Japanese phonology: "rendaku can be and should be accounted for in terms of phonological grammar."

5.6 Chapter summary

In this chapter, I have proposed a grammar model to account for the rendaku application patterns in Japanese surnames. I have discussed the challenge of modeling the lexicalized and productive aspects of rendaku in surnames with a single grammar. In order to address the issue, I have proposed a Maximum Entropy Harmonic Grammar model (see Goldwater and Johnson 2003) composed of general constraints and lexically-specific constraints (Moore-Cantwell and Pater 2016). Through learning simulations, I have shown that the proposed model with appropriate learning biases on constraint weights can not only capture the lexicalized rendaku patterns of existing surnames but also predict productive rendaku application in non-existing surnames. I have also argued that lexical factors should be incorporated into phonological grammar rather than simply specified in the lexicon.

CHAPTER 6

Conclusion

In this dissertation, I have investigated the phonological patterns of Japanese compound surnames with a focus on rendaku and accent. In this chapter, I conclude the thesis by giving brief summaries of the preceding chapters and discussing implications of the research for the field of phonology.

6.1 Summary of the dissertation

In Chapter 2, I have shown that the rendaku and prosodic patterns of compound surnames are different from those of common noun compounds. Their main characteristics include the effects of Strong Lyman's Law, whereby a voiced obstruent in the first element inhibits rendaku application, and an inverse correlation between the presence of accent and the presence of rendaku voicing. Many of these peculiarities have been noted in the literature (see Sugito 1965; Kubozono 2005; Zamma 2005 among others), but no previous study has ever given a principled explanation of the patterns.

In Chapter 3, I have addressed the issues by giving a formal analysis. I have argued that compound surnames are represented as single stems in the grammar and that rendaku application and accent assignment in surnames are regulated by the phonology of stems. I have shown that the proposed hypothesis can give a unified account of the peculiarities of surnames discussed in Chapter 2. The hypothesis based on stem phonology has made further predictions about how rendaku should apply in the kinds of surnames which have not been well described in the literature.

In Chapter 4, I have presented the results of a corpus study and an experiment testing the predictions of the proposed hyopthesis. The data of existing surnames collected from social media

have shown that rendaku application in surnames generally follows stem-internal phonological restrictions, supporting the hypothesis. The experimental results further show that Japanese speakers apply rendaku to nonce surnames in a similar way, providing evidence for the productivity of the patterns.

In Chapter 5, I have proposed a grammar model within the framework of Maximum Entropy Harmonic Grammar (see Goldwater and Johnson 2003). I have combined the MaxEnt grammar with general (morpho)phonological constraints and lexically-specific constraints which are indexed with every second element morpheme and with every compound surname along the lines of Moore-Cantwell and Pater (2016). I have shown that, with appropriate biases imposed on the learning of constraint weights, the model can account for both the lexicalized and productive aspects of rendaku application in surnames.

6.2 Implications for phonological theory

The findings of the study have various implications for the field of phonology. First, it opens new ways to investigate the sound patterns of proper nouns. It has been observed in the literature that names tend to show peculiar phonological behaviors as compared to common nouns. (See, for example, stress assignment in Turkish place names; Sezer 1981; Inkelas et al. 1996.) However, they are often deemed simply as exceptions or non-productive patterns. If any description is given, there is still no explanation as to *why* proper nouns show such peculiarities. As we have seen, rendaku and accent in compound surnames in Japanese is another example. Previous studies have noted their unique patterns, but no attempt has been made to explain the differences between surnames and regular words. This study has proven the productivity of the phenomenon with a nonce name experiment, and has given a formal analysis of the patterns treating compound surnames as single stems rather than as belonging to an exceptional category of stem-stem compounds. The results suggest that some, if not all, phonological phenomena found in proper nouns are productive and that they can be explained in terms of regular phonological grammar.

Second, the data of Japanese surnames and the analysis proposed here suggest that the repre-

sentations of "compound words" in language are not always uniform. As discussed above (Section 3.2.1), Hay (2003) has claimed that the pronunciations of words may affect their morphological representations and semantic properties. I have shown that the semantic properties of words may also affect their morphological representations, which in turn affect their phonological patterns. Other studies also report that morphologically-complex words may show properties of simplex words or vice versa, suggesting that the distinction between the two is not necessarily clear-cut (see e.g. Zuraw 2000; Zuraw and Peperkamp 2015). Japanese compound surnames present an extreme case where compounds always behave like single stems due to their semantic non-compositionality. The proposal about the representations of compound surnames provides a new analytical possibility for similar cases.[1]

Lastly, the overall results of the corpus study, the experiment, and the learning simulations presented in Chapter 4 and Chapter 5 show that lexicalization and productivity may coexist and that a single grammar can capture both at the same time. As discussed in Section 5.5, this provides the answer to the long-standing question of the status of rendaku in Japanese phonology (see Kawahara 2015a); despite its lexical irregularities, rendaku application is a productive process, which should be and can be addressed by phonological theory. Many of the phonological phenomena in languages actually involve lexically-conditioned patterns to a greater or lesser extent (see Moore-Cantwell and Pater 2016). As has been shown in Chapter 5, a MaxEnt HG model composed of general and lexically-specific constraints with learning biases enables us to account for both the lexicalized and productive aspects of phonological alternations, and also assess the balance between lexical factors and phonological factors involved in them. It is hoped that the current study will initiate further investigations into phonological phenomena involving lexicalization and productivity from both experimental and theoretical perspectives.

[1]Sino-Japanese compounds are also reported to behave like single stems and show what could be described as Strong Lyman's Law effects in rendaku application. See e.g. Vance (1996).

BIBLIOGRAPHY

Akinaga, Kazue. 1977a. Hime-kō: Rendaku o megutte [Thoughts on "hime": In relation to rendaku]. [Republished in Akinaga, K. (2009) Nihon-go on'in-shi, akusento shiron, Sakuma Shoin: Tokyo].

Akinaga, Kazue. 1977b. Matsubara to Yanagihara — ha-gyō tenko o chūshin ni [Matsubara and Yanagihara — with a special focus on the historical change of /h/]. *Kokugogaku* 111:62–76.

Alderete, John. 2015. Updating the analysis of Japanese compound accent. In Short 'chrift for Alan Prince compiled by Eric Baković; available at https://princeshortschrift.wordpress.com/.

Asai, Atsushi. 2014. Rendaku seiki no keikō to teichaku-ka [A tendency in the occurrence of rendaku and lexicalization]. *NINJAL Research Papers* 7:27–44.

Bates, Douglas, Martin Maechler, and Ben Bolker. 2011. lme4: Linear mixed-effects models using S4 classes. R package.

Becker, Michael, and Levine Jonathan. 2013. Experigen — an online experiment platform. Available at: http://becker.phonologist.org/experigen.

Berko, Jean. 1958. The child's learning of English phonology. *Word* 14:150–177.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row.

Endo, Kunimoto. 1980. Hi-rendaku no hōsoku no shōchō to sono imi: Daku-shiin to bion to no kankei kara [A decay of the rendaku-blocking law and its consequence: An examination through the relationship between voiced obstruents and nasals]. *Kokugo Kokubun* 50:38–54.

Frellesvig, Bijarke. 2010. *A history of the Japanese language*. Cambridge: Cambridge University Press.

Fukuda, Suzy, and Shinji Fukuda. 1994. To voice or not to voice: The operation of rendaku in Japanese developmentally language-impaired. *McGill Working Papers in Linguistics* 10:178–193.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 111–120. Stockholm: Stockholm University.

Hamada, Atsushi. 1952. Hatsuon to dakuon to no sōkan-sei no mondai [Issues in the relation between pronunciations and voiced obstruents]. *Kokugo Kokubun* 21:18–32.

Haraguchi, Shosuke. 2002. A theory of voicing. In *A comprehensive study on the phonological structure of languages and phonological theory*, ed. Shosuke Haraguchi, Technical Report of Basic Sciences (A)(1), Grant-in-Aid for Scientific Research by the Japan Society for the Promotion of Science, 1–22.

Hashimoto, Shinkichi. 1938/1950. Kokugo onin no hensen [Changes in the phonology of the national language]. In *Kokugo onin no kenkyū*, Hashimoto Shinkichi hakushi chosaku-shū 4, 51–103. Tokyo: Iwanami Shoten. [Originally published in 1938 in *Kokubungaku* 15(10): 3–41].

Hay, Jennifer. 2003. *Causes and consequences of word structure*. [Revised version of the author's Ph.D. dissertation (2000). Northwestern University]. New York: Routledge.

Hay, Jennifer, Janet Pierrehumbert, and Mary Beckman. 2004. Speech perception, well-formedness, and the statistics of the lexicon. In *Papers in Laboratory Phonology VI: Phonetic interpretation*, ed. John Local, Richard Ogden, and Rosalind Temple, 58–74. Cambridge: Cambridge University Press.

Hayashi, Ryosho. 1953a. Shomin no myōji [Family names of commoners]. *Nihon rekishi* 58:46–47.

Hayashi, Ryosho. 1953b. Shomin no myōji [Family names of commoners]. *Nihon rekishi* 62:49–50.

Hayes, Bruce, and Tanya Stivers. 1996. The phonetics of post-nasal voicing. Ms. UCLA.

Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379–440.

Hayes, Bruce, Kie Zuraw, Peter Siptár, and Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85:822–863.

Hepburn, James Curtis, ed. 1867/1872. *Waei gorin shūsei [Japanese-English and English-Japanese dictionary]*. Shanghai: The American Presbyterian Mission Press, 2nd edition.

Hirano, Takanori. 2013. A rule application approach to rendaku. Paper presented at ICPP 2013 Tokyo.

Hirata, Junko. 2010. Rendaku shiron [An essay on rendaku]. *Kobe Kaisei Joshi Gakuin Daigaku Kenkyū Kiyō* 49:31–37.

Hirata, Junko. 2011. Rendaku shiron (sono 2) [An essay on rendaku 2]. *Kobe Kaisei Joshi Gakuin Daigaku Kenkyū Kiyō* 50:89–93.

Hora, Tomio. 1952. Edo-jidai-no shomin-wa hatashite myōji-o motanakatta-ka [Did commoners in the Edo period really not have surnames?]. *Nihon rekishi* 50-7:2–7.

Hora, Tomio. 1966. *Shomin kazoku no rekishi-zō [A historical image of ordinary families]*. Tokyo: Koso Shobo.

Ihara, Mutsuko, and Tadao Murata. 2006. Nihon-go no rendaku ni kansuru ikutsuka no jikken [Experiments on rendaku in Japanese]. *On'in Kenkyū* 9:17–24.

Ihara, Mutsuko, Katsuo Tamaoka, and Tadao Murata. 2009. Lyman's Law effect in Japanese sequential voicing: Questionnaire-based nonword experiments. In *Current issues in unity and diversity of languages: Collection of the papers selected from the 18th International Congress of Linguists*, ed. The Linguistic Society of Korea, 1007–1018. Seoul: Dongam Publishing.

Inkelas, Sharon, Cemil Orhan Orgun, and Chryl Zoll. 1996. Exceptions and static phonological patterns: Cophonologies vs. pre-specification. ROA-124-0496.

Irwin, Mark. 2005. Rendaku-based lexical hierarchies in Japanese: The behaviour of Sino-Japanese mononoms in hybrid noun compounds. *Journal of East Asian Linguistics* 14:121–153.

Irwin, Mark. 2009. Prosodic size and rendaku immunity. *Journal of East Asian Linguistics* 18:179–196.

Irwin, Mark. 2012. Rendaku dampening and prefixes. *NINJAL Research Papers* 4:27–36.

Irwin, Mark. 2014a. Rendaku across duplicate moras. *NINJAL Research Papers* 7:93–109.

Irwin, Mark. 2014b. Rendaku lovers, rendaku haters and the logistic curve. In *Japanese/Korean Linguistics*, ed. Mikio Giriko, Naonori Nagaya, Akiko Takemura, and J. Timothy Vance, volume 22, 37–51. Stanford: CSLI Publications.

Irwin, Mark. 2016a. The rendaku database. In *Sequential voicing in Japanese: Papers from the NINJAL Rendaku Project*, ed. Timothy J. Vance and Mark Irwin, number 176 in Studies in Language Companion, chapter 6, 79–106. Amsterdam and Philadelphia: John Benjamins.

Irwin, Mark. 2016b. Rosen's Rule. In *Sequential voicing in Japanese: Papers from the NINJAL Rendaku Project*, ed. Timothy J. Vance and Mark Irwin, number 176 in Studies in Language Companion, chapter 7, 107–117. Amsterdam and Philadelphia: John Benjamins.

Irwin, Mark, and Mizuki Miyashita. 2013-2016. The rendaku database version 2.0-2.8. Available online at:
http://www-h.yamagata-u.ac.jp/˜irwin/site/Rendaku_Database.html.

Irwin, Mark, Mizuki Miyashita, and Kerri Russell. 2017. The rendaku database version 3.1. Available online at:
http://www-h.yamagata-u.ac.jp/˜irwin/site/Rendaku_Database.html.

Ishizuka, Tatsumaro. 1801. *Kogen seidaku kō [An account of voicing in Old Japanese]*. Kyoto: Hishiya Magobei, Zeniya Rihee, Hayashi Ihee and Kashiwaya Heisuke.

Ito, Junko. 1986. Syllable theory in prosodic phonology. Doctoral Dissertation, University of Massachusetts Amherst, Amherst.

Ito, Junko, and Armin Mester. 1986. The phonology of voicing in Japanese: Theoretical consequences for morphological accessibility. *Linguistic Inquiry* 17:49–73.

Ito, Junko, and Armin Mester. 1992/2003. Weak layering and word binarity. In *A new century of phonology and phonological theory: A festschrift for Professor Shosuke Haraguchi on the occasion of his sixtieth birthday*, ed. Takeru Homma, Masao Okazaki, Toshiyuki Tabata, and Shin-ichi Tanaka, 26–65. Tokyo: Kaitakusha. Originally published as Linguistic Research Center Working Paper LRC-92-09. UC Santa Cruz.

Ito, Junko, and Armin Mester. 1994. Reflections on CodaCond and Alignment. In *Phonology at Santa Cruz*, volume 3, 27–46. Linguistics Research Center, University of California, Santa Cruz.

Ito, Junko, and Armin Mester. 1995a. The core-periphery structure of the lexicon and constraints on reranking. In *University of Massachusetts occasional papers in linguistics [UMOP] 18: Papers in Optimality Theory*, ed. Jill Beckman, Suzanne Urbanczyk, and Laura Walsh Dickey, 181–209. Amherst: GLSA.

Ito, Junko, and Armin Mester. 1995b. Japanese phonology: Constraint domains and structure preservation. In *The handbook of phonological theory*, ed. John Goldsmith, 817–838. Cambridge, MA: Blackwell Publishers.

Ito, Junko, and Armin Mester. 1996. Stem and word in Sino-Japanese. In *Phonological structure and language processing: Cross-linguistic studies*, ed. Takashi Otake and Anne Cutler, 13–44. Berlin: Mouton de Gruyter.

Ito, Junko, and Armin Mester. 1998. Markedness and word structure: OCP effects in Japanese. Manuscript. University of California, Santa Cruz. [Available on Rutgers Optimality Archive, http://roa.rutgers.edu, ROA-255-0498.].

Ito, Junko, and Armin Mester. 1999. The phonological lexicon. In *The handbook of Japanese linguistics*, ed. Natsuko Tsujimura, 62–100. Oxford: Blackwell Publishers.

Ito, Junko, and Armin Mester. 2003. *Japanese morphophonemics*. Cambridge, MA: MIT Press.

Ito, Junko, and Armin Mester. 2015. Sino-Japanese phonology. In *The handbook of Japanese language and linguistics: Phonetics and phonology*, ed. Haruo Kubozono, 289–312. Berlin: De Gruyter Mouton.

Ito, Junko, and Armin Mester. 2016. Unaccentedness in Japanese. *Linguistic Inquiry* 47:3:471–526.

Iwasaki, Shoichi. 2013. *Japanese: Revised edition*. London Oriental and African Language Library 17. Amsterdam and Philadelphia: John Benjamins.

Jäger, Gerhard. 2007. Maximum entropy models and stochastic optimality theory. In *Architectures, rules, and preferences: Variations on themes by Joan Bresnan*, ed. Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling, and Chris Manning, 467–479. Stanford: CSLI Publications.

Jesuit missionaries, ed. 1603/1604. *Nippo jisho [Japanese-Portuguese Didctionary]; [original title: Vocabulario da lingoa de Iapam com adeclaração em Portugues]*. Nagasaki: Nagasaki Gakurin.

Jodaigo Jiten Henshu Iinkai, Hisataka Omodaka, Toru Asami, Teizo Ikegami, Itaru Ide, Ito Haku, Kawabata Yoshiaki, Masatoshi Kinoshita, Noriyuki Kojima, Atsuyoshi Sakakura, Akihiro Satake, Kazutami Nishimiya, and Shiro Hashimoto, ed. 1967. *Jidaibetsu kokugo daijiten Jōdaihen [Unabridged dictionary of the national language by age: Old Japanese]*. Tokyo: Sanseido.

Kamo, no Mabuchi. 1765-1789. Goi-kō [An account of word meanings]. [Republished in Kamo no Mabuchi Zenshū [The Complete Works of Kamo no Mabuchi] (1977-1992). Zokugun Shorui-jū Kansei-kai, Tokyo].

Kato, Akira. 1984. Nihon no seishi [Surnames in Japan]. In *Higashi asia sekai ni okeru Nihon kodaishi kōza*, ed. Mitsusada Inoue, Sadao Nishijima, Yukio Takeda, and Ken Amakasu. Tokyo: Gakuseisha.

Kawahara, Shigeto. 2002. Similarity among variants: Output-variant correspondence. Bachelor's thesis, International Christian University.

Kawahara, Shigeto. 2006. A faithfulness ranking projected from a perceptibility scale: The case of voicing in Japanese. *Language* 82:536–574.

Kawahara, Shigeto. 2008. Phonetic naturalness and unnaturalness in Japanese loanword phonology. *Journal of East Asian Linguistics* 18:317–330.

Kawahara, Shigeto. 2011. Japanese loanword devoicing revisted: A rating study. *Natural Language and Linguistic Theory* 29:705–723.

Kawahara, Shigeto. 2012. Lyman's Law is active in loanwords and nonce words: Evidence from naturalness judgment studies. *Lingua* 122:1193–1206.

Kawahara, Shigeto. 2015a. Can we use rendaku for phonological argumentation? Linguistics Vanguard, doi:10.1515/lingvan-2015-0001.

Kawahara, Shigeto. 2015b. The phonology of Japanese accent. In *The handbook of Japanese language and linguistics: Phonetics and phonology*, ed. Haruo Kubozono, 445–492. Berlin: De Gruyter Mouton.

Kawahara, Shigeto. 2016. Psycholinguistic studies of rendaku. In *Sequential voicing in Japanese: Papers from the NINJAL Rendaku Project*, ed. Timothy J. Vance and Mark Irwin, number 176 in Studies in Language Companion, chapter 3, 35–45. Amsterdam and Philadelphia: John Benjamins.

Kawahara, Shigeto, Hajime Ono, and Kiyoshi Sudo. 2006. Consonant co-occurrence restrictions in Yamato Japanese. In *Japanese/Korean Linguistics*, ed. Timothy J. Vance and Kimberley Jones, volume 14, 27–38. Stanford: CSLI Publications.

Kawahara, Shigeto, and Shin-ichiro Sano. 2014a. Identity avoidance and Lyman's Law. *Lingua* 150:71–77.

Kawahara, Shigeto, and Shin-ichiro Sano. 2014b. Identity avoidance and rendaku. In *Proceedings of Phonology 2013*, ed. John Kingston, Claire Moore-Cantwell, Joe Pater, and Robert Staubs.

Kawahara, Shigeto, and Shin-ichiro Sano. 2014c. Testing Rosen's Rule and Strong Lyman's Law. *NINJAL Research Papers* 7:111–120.

Kawahara, Shigeto, and Shin-ichiro Sano. 2016. Rendaku and identity avoidance: Consonantal identity and moraic identity. In *Sequential voicing in Japanese: Papers from the NINJAL Rendaku Project*, ed. Timothy J. Vance and Mark Irwin, number 176 in Studies in Language Companion, chapter 4, 47–55. Amsterdam and Philadelphia: John Benjamins.

Kindaichi, Haruhiko. 1965. *Kotoba no saijiki*. Tokyo: Shinchosha.

Kindaichi, Haruhiko. 1976/2005. Rendaku no kai [An account of rendaku]. In *Kindaichi Haruhiko chosakushū*, volume 6, 583–614. Tokyo: Tamagawa Daigaku Shuppanbu. [Originally published in 1976 in *Sophia Linguistica* 2:1–22].

Kindaichi, Haruhiko, Ōki Hayashi, and Takeshi Shibata, ed. 1988. *Nihongo hyakka daijiten [An encyclopaedia of the Japanese language]*. Tokyo: Taishūkan.

Kindaichi, Kyosuke. 1938. *Kokugo oninron [The phonology of the national language]*. Toko Shoin.

Kiyose, Gisaburo. 1985. Heianchō hagyō-shiin p-onron. *Onsei no Kenkyū* 21:73–87.

Kubozono, Haruo. 1995. Constraint interaction in Japanese phonology: Evidence from compound accent. In *Phonology at Santa Cruz [PASC]*, ed. Rachel Walker, Ove Lorentz, and Haruo Kubozono, 21–38. Santa Cruz: Santa Cruz: Linguistics Research Center.

Kubozono, Haruo. 1997. Lexical markedness and variation: A nonderivational account of Japanese compound accent. In *Proceedings of the West Coast Conference on Formal Linguistics 15*, volume 15, 273–287.

Kubozono, Haruo. 2005. Rendaku: Its domain and linguistic conditions. In *Voicing in Japanese*, ed. Jeroen van de Weijer, Kensuke Nanjo, and Tetsuo Nishihara, 5–24. Berlin and New York: Mouton de Gruyter.

Kubozono, Haruo. 2006. Where does loanword prosody come from? A case study of Japanese loanword accent. *Lingua* 116:1140–1170.

Kubozono, Haruo. 2008. Japanese accent. In *The Oxford handbook of Japanese linguistics*, ed. Shigeru Miyagawa and Mamoru Saito, 165–191. Oxford: Oxford University Press.

Kubozono, Haruo. 2015. Introduction to Japanese phonetics and phonology. In *The handbook of Japanese phonetics and phonology*, ed. Haruo Kubozono, Handbooks of Japanese Language and Linguistics 2, 1–40. Berlin and New York: Mouton de Gruyter.

Kuginuki, Toru. 1982. Jōdai Nihon-go ra-gyōon-kō [A study on /r/ in Old Japanese]. *Toyama Daigaku Jinbungakubu Kiyō* 6:192–206.

Kumagai, Gakuji. 2016. The ganging-up of OCP-labial effect on Japanese rendaku. Paper presented at the 2016 Annual Meeting on Phonology (AMP).

Kumagai, Gakuji. 2017. Testing OCP-labial effect on Japanese rendaku. Lingbuzz/003290.

Kuznetsova, Alexandra, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. 2013. lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R package.

Labrune, Laurence. 2012. *The Phonology of Japanese (The Phonology of the World's Languages)*. Oxford: Oxford University Press.

Labrune, Laurence. 2014. The phonology of Japanese /r/: A panchronic account. *Journal of East Asian Linguistics* 23:1–25.

Labrune, Laurence. 2016. Rendaku in cross-linguistic perspective. In *Sequential voicing in Japanese: Papers from the NINJAL Rendaku Project*, ed. Timothy J. Vance and Mark Irwin,

number 176 in Studies in Language Companion, chapter 8, 195–233. Amsterdam and Philadelphia: John Benjamins.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 1990. Harmonic grammar — A formal multi-level connectionist theory of linguistic well-formedness: An application. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 884–891. Mahwah, NJ: Lawrence Erlbaum Associates.

Lyman, Benjamin. 1885. On the Japanese nigori of composition. *Journal of the American Oriental Society* 11:142–143.

Lyman, Benjamin. 1894. The change from surd to sonant in Japanese compounds. In *Oriental studies: A selection of the papers read before the Oriental Club in Philadelphia 1888-1894*, ed. The Oriental Club in Philadelphia, 160–176. Boston: Ginn and company.

Martin, Andrew. 2007. The evolving lexicon. Doctoral Dissertation, University of California, Los Angeles.

Martin, Andrew. 2011. Grammars leak: Modeling how phonotactic generalizations interact within the grammar. *Language* 87:751–770.

Martin, Samuel E. 1952. *Morphophonemics of Standard Colloquical Japanese*. Number 47 in Language dissertation. Baltimore: The Linguistic Society of America.

Martin, Samuel E. 1987. *The Japanese language through time*. Yale University Press.

Matsuura, Yoko. 1996. Jiongo no rendaku ni okeru senkō onsetsu no eikyō ni tsuite [The effect of the preceding syllable on the 'rendaku' in Sino-Japanese words]. *Hiroshima Daigaku Nihongo Kyōiku Hiroshimaō* 6:37–43.

McCarthy, John. 2003. OT constraints are categorical. *Phonology* 20:75–138.

McCarthy, John. 2008. *Doing Optimality Theory*. Malden: Blackwell Publishing.

McCarthy, John, and Alan Prince. 1993. Generalized alignment. In *Yearbook of Morphology (1993)*, ed. Geert Booij and Jaap van Marle, 79–153. Springer.

McCarthy, John, and Alan Prince. 1995. Faithfulness and reduplicative identity. In *University of Massachusetts occasional papers in linguistics [UMOP] 18: Papers in Optimality Theory*, ed. Jill Beckman, Suzanne Urbanczyk, and Laura Walsh Dickey, 249–384. Amherst: GLSA.

McCawley, James. 1968. *The phonological component of a grammar of Japanese*. [Revised version of the author's Ph.D. dissertation (1965). *The accentual system of Standard Japanese*. Massachusetts Institute of Technology]. The Hague: Mouton.

Meiji Government, The. 1870. *Heimin myōji kyoka rei [Act on the permission of the use of surnames by commoners]*. Hōrei Zensho. [Published by the Meiji Cabinet in 1887]. Available in the National Diet Library Digital Collections: http://dl.ndl.go.jp/info:ndljp/pid/787950/212. Tokyo: Naikaku Kanpō-kyoku.

Meiji Government, The. 1875. *Heimin myōji hisshō gimu rei [Act on the obligatory use of surnames by commoners]*. Hōrei Zensho. [Published by the Meiji Cabinet in 1887]. Available in the National Diet Library Digital Collections: http://dl.ndl.go.jp/info:ndljp/pid/787955/71. Tokyo: Naikaku Kanpō-kyoku.

Miller, George A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63:81–97.

Miura, Keiichi. 2012. *Yamaha sōsōfu [The history of Yamaha]*. Hamamatsu: Ankasha.

Miyake, Hideo Marc. 2003. *Old Japanese: A phonetic reconstruction*. London: Routledge.

Miyake, Takeo. 1932. Dakuonkō [An examination of voiced obstruents]. *Onsei no kenkyū* 5:135–190.

Miyashita, Mizuki, Mark Irwin, and Timothy J. Vance. 2016. Rendaku in tōhoku Japanese: The Kahoku-chō survey. In *Sequential voicing in Japanese: Papers from the NINJAL Rendaku*

Project, ed. Timothy J. Vance and Mark Irwin, number 176 in Studies in Language Companion, chapter 10, 173–193. Amsterdam and Philadelphia: John Benjamins.

Moore-Cantwell, Claire. 2015. The phonological grammar is probabilistic: New evidence pitting abstract representation against analogy. Unpublished manuscript, Yale University. Available at: http://blogs.ubc.ca/amp2015/files/2015/09/Moore- Cantwell.pdf.

Moore-Cantwell, Claire, and Joe Pater. 2016. Gradient exceptionality in Maximum Entropy Grammar with lexically specific constraints. *Catalan Journal of Linguistics* 15:53–66.

Morioka, Hiroshi. 2011. *Myōji no nazo [The mysteries of surnames]*. Tokyo: Chikuma Shobo.

Morita, Takeshi. 1977. Nippo-jisho ni mieru goon-ketsugō-jō no ichi-keikō [A phonological tendency in compounding found in Japanese-Portuguese dictionaries]. *Kokugogaku* 108:20–29.

Motoori, Norinaga. 1790-1822. *Kojiki-den [Commentaries on the Kojiki, Records of Acient Matters]*. Nagoya: Eirakuya.

Murayama, Shichiro. 1954. Rendaku ni tsuite [On sequential voicing]. *Gengo Kenkyū* 26-27:106–110.

Nakagawa, Yoshio. 1966. Rendaku, rensei (kashō) no keifu [A genealogy of sequential voicing and sequential non-voicing (working label)]. *Kokugo Kokubun* 35-6:302–314.

Nakagawa, Yoshio. 1978. Koyū meishi no rendaku rensei no keifu [A genealogy of sequential voicing and sequential non-voicing in proper nouns]. *Shizuoka Joshi Daigaku Kokubun Kenkyū* 12:288–302.

National Institute for Japanese Language, The. 2012. The corpus of spontaneous Japanese. http://pj.ninjal.ac.jp/corpus_center/csj/misc/preliminary/index_e.html.

Nihon Kokugo Daijiten Dainihan Henshū Iinkai, ed. 2000. *Nihon kokugo daijiten [Japanese dictionary]*. Tokyo: Shogakukan, 2nd edition.

Nishimura, Kohei. 2003. Lyman's law in loanwords. Master's thesis, Nagoya University.

Nishiyama, Kunio. 2010. Penultimate accent in Japanese predicates and the verb-noun distinction. *Lingua* 120:2353–2366.

Ohno, Kazutoshi. 2000. The lexical nature of rendaku in Japanese. In *Japanese/Korean Linguistics*, ed. Mineharu Nakayama and Charles Quinn, volume 9, 151–164. Stanford: CSLI Publications.

Ohta, Satoshi. 2013. On the relationship between rendaku and accent: Evidence from -kawa/-gawa alternation in Japanese surnames. In *Current issues in Japanese phonology: Segmental variation in Japanese*, ed. Jeroen van de Weijer and Tetsuo Nishihara, 63–87. Tokyo: Kaitakusha.

Okumura, Mitsuo. 1952. Jion no shindaku ni tsuite [On new voicing in Sino-Japanese words]. *Kokugo Kokubun* 21:327–340.

Okumura, Mitsuo. 1984. Rendaku. *Nihongogaku* 3:89–98.

Okutomi, Takayuki. 2004. *Myōji no rekishi-gaku [A historical study of surnames]*. Tokyo: Kadokawa Shoten.

Otsu, Yukio. 1980. Some aspects of rendaku in japanese and related problems. In *Theoretical issues in Japanese linguistics (MIT Working Papaers in Linguistics 2)*, ed. Yukio Otsu and Anne Farmer, 207–227. Cambridge: MIT Press.

Pater, Joe. 1999. Austronesian nasal substitution and other nc effects. In *The prosody-morphology interface*, ed. René Kager, Harry van der Hulst, and Wim Zonneveld, 310–343. Cambridge University Press.

Pater, Joe. 2007. The locus of exceptionality: Morpheme-specific phonology as constraint indexation. In *Papers in Optimality Theory III*, ed. Leah Bateman, Michael O'Keefe, Ehren Reilly, and Adam Werle, University of Massachusetts Occasional Papers in Linguistics 32, 259–296. Amherst: University of Massachusetts, Graduate Linguistic Student Association.

Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 36:999–1035.

Poser, William. 1984. The phonetics and phonology of tone and intonation in Japanese. Doctoral Dissertation, Massachussetts Institute of Technology.

Prince, Alan, and Paul Smolensky. 1993/2004. *Optimality Theory: Constraint interaction in generative grammar*. Cambridge, MA: Blackwell Publishers.

R Development Core Team. 1993-2017. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Ramsey, Robert, and Marshall Unger. 1972. Evidence for a consonant shift in 7th century Japanese. *Papers in Japanese Linguistics* 1:279–295.

Rice, Keren. 1997. Japanese NC clusters and the reduncancy of postnasal voicing. *Linguistic Inquiry* 28:541–551.

Rosen, Eric. 2001. Phonological processes interacting with the lexicon: Variable and non-regular effects in Japanese phonology. Doctoral Dissertation, University of British Columbia.

Rosen, Eric. 2003. Systematic irregularity in Japanese rendaku: How the grammar mediates patterned lexical exceptions. *Canadian Journal of Linguistics* 48:1–37.

Rosen, Eric. 2016. Predicting the unpredictable: capturing the apparent semi-regularity of rendaku voicing in Japanese through Gradient Symbolic Computation. In *Proceedings of the 42nd Annual Meeting of the Berkeley Linguistics Society*, ed. Emily Clem, Geoff Bacon, Andrew Cheng, Virginia Dawson, Erik Maier Hans, Alice Shen, and Amalia Horan Skilton. Berkeley: Berkeley Linguistics Society.

Saffran, Jenny R., Richard J. Aslin, and Elissa L. Newport. 1996a. Statistical learning by 8-month old infants. *Science* 274:1926–1928.

Saffran, Jenny R., Elissa L. Newport, and Richard J. Aslin. 1996b. Word segmentation: The role of distributional cues. *Journnal of Memory and Language* 35:606–621.

Sakata, Satoshi. 2006. *Myōji to namae no rekishi [History of surnames and given names]*. Rekishi Bunka Library. Tokyo: Yoshikawa Kobunkan.

Sakurai, Shigeharu. 1972. Heian Insei jidai ni okeru wago no rendaku ni tsuite [On rendaku in native words in Japanese of the Heian Insei era]. *Kokugo Kokubun* 41:1–19.

Sano, Shin-ichiro. 2015. Universal markedness reflected in the patterns of voicing process. In *Proceedings of the North East Linguistic Society (NELS)*, ed. Thuy Bui and Deniz Özyıldız, volume 45-3, 49–58.

Sato, Hirokazu. 1989. Fukugō-go ni okeru akusento kisoku to rendaku kisoku [Accent rules and rendaku rules in compounds]. In *Kōza nihongo to nihongo kyōiku 2: Nihongo no onsei, on'in (jō) [Japanese and Japanese teaching 2: Japanese phonetics, phonology 1]*, 233–265. Tokyo: Meiji Shoin.

Sezer, Engin. 1981. The k/0 alternation in Turkish. In *Harvard studies in phonology*, ed. Nick Clements, 354–382. Bloomington: Indiana University Linguistics Club.

Shinmura, Izuru, ed. 1998. *Kōjien*. Japanese dictionary. Tokyo: Iwanami Shoten, 2nd edition.

Shinmura, Izuru, ed. 2008. *Kōjien*. Japanese dictionary. Tokyo: Iwanami Shoten, 6th edition.

Shinohara, Shigeko. 2000. Default accentuation and foot structure in Japanese: Evidence from Japanese adaptation of French words. *Journal of East Asian Linguistics* 9:55–96.

Shirooka, Keiji, and Tadashige Murayama. 2011. A database of Japanese surnames and their rankings. Available on-line at:
http://www.ipc.shizuoka.ac.jp/˜jjksiro/kensaku.html.

Smolensky, Paul, and Matthew Goldrick. 2015. Gradient Symbolic Computation. LSA Summer Institute Workshop.

Sugawara, Ayaka. 2012. Japanese accent is largely predictable: Evidence from given names. Slides for the 144th meeting of the Linguistic Society of Japan, Tokyo University of Foreign Studies, Tokyo, Japan.

Sugito, Miyoko. 1965. Shibata-san to Imada-san: Tango no chōkakuteki benbetsu ni tsuite no ichi kōsatsu [Shibata-san and Imada-san: An examination of auditory distinction of words]. *Gengo Seikatsu* 165:64–72.

Suzaki, Haruo. 2013. A private on-line corpus of 111,711 Japanese family names. http://www2s.biglobe.ne.jp/˜suzakihp/index40.html.

Suzuki, Hiroaki. 2016. *Kyōyō to shite no ninchi-kagaku [Cognitive science as general education]*. Tokyo: University of Tokyo Press.

Suzuki, Yutaka. 2004. "Rendaku" no koshō ga kakuritsu suru made: Rendaku kenkyū zenshi [The origin of the term "rendaku": The background history of rendaku]. *Kokubungaku Kenkyu* 142:124–134.

Suzuki, Yutaka. 2005. Lyman no hōsoku no reigai ni tsuite: Rendaku-kei "-basigo" o kōbu-seiso to suru fukugō-go o chūshin ni [On exceptions to Lyman's Law: Compounds with the rendaku form "-basigo" as the second element]. *Bunkyo Gakuin Daigaku Gaikokugo Gakubu Bunkyo Gakuin Tanki Daigaku Kiyō* 4:249–265.

Suzuki, Yutaka. 2008. Lyman no hōsoku reigai no seiritsu-katei ni tsuite: "takara-gai" o kōbu-yōso to suru go no rendaku [On the origin of exceptions to Lyman's Law: Rendaku in compounds with "takara-gai" as the second element]. *Bunkyo Gakuin Daigaku Gaikokugo Gakubu Bunkyo Gakuin Tanki Daigaku Kiyō* 7:279–294.

Suzuki, Yutaka. 2015. Hime-kō zokuchō: "Kojiki" ni okeru /-hime/ to /-hiko/ no rendaku. *Akusento-shi shiryō kenkyū-kai ronshū X (Akinaga Kazue sensei beiju kinen)* 10:11–30.

Takayama, Michiaki. 1992. Rendaku to renjōdaku [On sequential voicing and sequential post-nasal voicing]. *Kuntengo to Kunten Shiryō* 88:115–124.

Takayama, Michiaki. 2012. *Nihon-go on'in-shi no kenkyū [A study on the history of Japanese phonology]*. Tokyo: Hitsuji Shobō.

Tamaoka, Katsuo, Mutsuko Ihara, Tadao Murata, and Hyunjung Lim. 2009. Effects of first-element phonological-length and etymological-type features on sequential voicing (rendaku) of second elements. *Journal of Japanese Linguistics* 25:17–38.

Tamura, Yoshinaga. 1953. Tokugawa jidai shomin no myōji [Family names of commoners in the Tokugawa period]. *Nihon rekishi* 60:314–315.

Tanaka, Shin-Ichi. 2001. The emergence of 'unaccented': Possible patterns and variations in Japanese compound accentuation. In *Issues in Japanese phonology and morphology*, ed. Jeroen van de Weijer and Tetsuo Nishihara, 159–192. Dordrecht: Mouton de Gruyter.

Tanaka, Shin-Ichi. 2005a. *Akusento to rizumu [Accent and rhythm]*. Tokyo: Kenkyusha.

Tanaka, Shin-Ichi. 2005b. Where voicing and accent meet: Their function, interaction, and opacity problems in phonological prominence. In *Voicing in Japanese*, ed. Jeroen van de Weijer, Kensuke Nanjo, and Tetsuo Nishihara, 261–278. Berlin and New York: Mouton de Gruyter.

Tanaka, Shin-ichi, and Haruo Kubozono. 1999. *Nihon-go no hatsuon kyōshitsu: Riron to renshū [A course in Japaense pronunciation: Theory and practice]*. Tokyo: Kuroshio Shuppan.

Tanaka, Yu. 2017. Phonotactically-driven rendaku in surnames: A linguistic study using social media. In *Proceedings of the West Coast Conference on Formal Linguistics 34*, ed. Aaron Kaplan, Abby Kaplan, Miranda K. McCarvel, and Edward J. Rubin, 519–528. Somerville, MA: Cascadilla Proceedings Project.

Tanaka, Yu, and Jun Yashima. 2013. Deliberate Markedness in Japanese hypocoristics. In *Proceedings of GLOW in Asia IX 2012*, ed. Nobu Goto, Koichi Otaki, Atsushi Sato, and Kensuke Takita, 283–297.

Tateishi, Koichi. 1989. Phonology of Sino-Japanese morphemes. *University of Massachusetts Occasional Papers in Linguistics* 13:209–235.

Toda, Ayako. 1988. Wago no hi-rendaku kisoku to rendaku keikō [The rendaku-blocking rules and rendaku-inhibiting factors in native Japanese words]. *Dōshisha Kokubungaku* 30:80–98.

Toyoda, Takeshi. 1971. *Myōji no rekishi [History of surnames]*. Chuo shinsho. Tokyo: Chuo Koronsha.

Toyota Motor Corporation. 1995-2017a. Webpage: "75 Years of TOYOTA" (last accessed in January 2017).
http://www.toyota-global.com/company/history_of_toyota/75years/text/.

Toyota Motor Corporation. 1995-2017b. Webpage: "The history of the emblem of Toyota Motor Corporation" (last accessed in January 2017).
http://www.toyota-global.com/showroom/emblem/history/.

Ueda, Kazutoshi. 1898. Gogaku sōken: p-on-kō [A new linguistic perspective: On the sound /p/]. *Teikoku Bungaku* 4:41–46.

Unger, J. Marshall. 2004. Alternations of *m* and *b* in Early Middle Japanese: The deeper signifiance of the sound-symbolic stratum. *Japanese Language and Literature* 38:323–337.

Unger, Marshall. 1977. *Studies in Early Japanese morphophonemics*. [Ph.D. dissertation (1975). Yale University. Bloomington: Indiana University Linguistics Club.

Vance, Timothy J. 1979. Nonsense-word experiments in phonology and their application to rendaku in Japanese. Doctoral Dissertation, University of Chicago.

Vance, Timothy J. 1980a. Comments on "Some aspects of rendaku in Japanese and related problems". In *Theoretical issues in Japanese linguistics (MIT Working Papaers in Linguistics 2)*, ed. Yukio Otsu and Anne Farmer, 229–236. Cambridge: MIT Press.

Vance, Timothy J. 1980b. The psychological status of a constraint on Japanese consonant alternation. *Linguistics* 18:245–267.

Vance, Timothy J. 1982. On the origin of voicing alternation in Japanese consonants. *Journal of the American Oriental Society* 102:333–341.

Vance, Timothy J. 1987. *An Introduction to Japanese Phonology*. Albany: State University of New York Press.

Vance, Timothy J. 1996. Sequential voicing in Sino-Japanese. *The Journal of the Association of Teachers of Japanese* 30:22–43.

Vance, Timothy J. 2005a. Nihongo kyōiku ni okeru rendaku [Rendaku in Japanese language education]. In *Gengogaku to Nihongo kyōiku [Linguistics and Japanese language education]*, ed. Masahiko Minami, volume 4, 1–11. Tokyo: Kuroshio Shuppan.

Vance, Timothy J. 2005b. Rendaku in inflected words. In *Voicing in Japanese*, ed. Jeroen van de Weijer, Kensuke Nanjo, and Tetsuo Nishihara, 89–103. Berlin and New York: Mouton de Gruyter.

Vance, Timothy J. 2005c. Sequential voicing and Lyman's Law in Old Japanese. In *Polymorphous linguistics: Jim McCawley's legacy*, ed. Salikoko S. Mufwene, Elaine J. Francis, and Rebecca S. Wheeler, 27–43. Cambridge, MA: MIT Press.

Vance, Timothy J. 2014. If rendaku isn't a rule, what in the world is it? In *Usage-based approaches to Japanese grammar: Towards the understanding of human language*, ed. Kaori Kabata and Tsuyoshi Ono, 137–152. Amsterdam: John Benjamins.

Vance, Timothy J. 2015. Rendaku. In *The handbook of Japanese phonetics and phonology*, ed. Haruo Kubozono, Handbooks of Japanese Language and Linguistics 2, 397–441. Berlin and New York: Mouton de Gruyter.

Vance, Timothy J., and Atsushi Asai. 2016. Rendaku and individual segments. In *Sequential voicing in Japanese: Papers from the NINJAL Rendaku Project*, ed. Timothy J. Vance and Mark Irwin, number 176 in Studies in Language Companion, chapter 8, 119–137. Amsterdam and Philadelphia: John Benjamins.

Vance, Timothy J., and Mark Irwin. 2013. A rendaku database for Old Japanese. Paper presented at the 21st International Conference on Historical Linguistics, University of Oslo.

Vance, Timothy J., Mizuki Miyashita, and Mark Irwin. 2014. Rendaku in Japanese dialects that

retain prenasalization. In *Japanese/Krean Linguistics*, volume 21, 33–42. Stanford: CSLI Publications.

Walter, Mary Ann. 2007. Repetition avoidance in human language. Doctoral Dissertation, Massachusetts Institute of Technology.

Watanabe, Toshiro, Edmund R. Skrzypczak, and Paul Snowden, ed. 2008. *Kenkyusha's new Japanese-English dictionary*. Tokyo: Kenkyusha, 5th edition.

Westbury, John, and Patricia Keating. 1986. On the naturalness of stop consonant voicing. *Journal of Linguistics* 22:145–166.

Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30:945–982.

Wilson, Colin. 2014. Maximum Entropy models. A tutorial at Annual Meeting on Phonology (AMP) 2014, MIT, September 19, 2014.

Wilson, Colin, Ben George, and Bruce Hayes. 2006. UCLA Maxent Grammar Tool. Available at: http://www.linguistics.ucla.edu/people/hayes/MaxentGrammarTool/.

Yamaguchi, Kyoko. 2011. Accentedness and rendaku in Japanese deverbal compounds. *Gengo Kenkyū* 140:117–133.

Yamaguchi, Yoshinori. 1988. Kodai-go no fukugō-go ni kansuru ichi-kōsatsu: Rendaku o megutte [a study of compounds in Old Japanese: On rendaku]. *Nihongogaku* 7-5:4–12.

Yanagita, Kunio. 1933/1998. Chimei sonota no hanashi [Stories on place names and others]. In *Yanagita Kunio zenshū*, volume 7. Tokyo: Chikuma Shobo. [Originally published in 1933].

Yanagita, Kunio. 1946/1998. Ie kandan [Episodes on families]. In *Yanagita kunio zenshū*, volume 15. Tokyo: Chikuma Shobo. [Originally published in 1946].

Yip, Moira. 1998. Identity avoidance in phonology and morphology. In *Morphology and its relation to phonology and syntax*, ed. Steven G. LaPointe, Diane K. Brentari, and Patrick M. Farrell, 216–246. Stanford: CSLI Publications.

Zamma, Hideki. 2001. Accentuation of person names in Japanese and its theoretical implications. *Tsukuba English Studies* 20:1–18.

Zamma, Hideki. 2005. The correlation between accentuation and rendaku in Japanese surnames: A morphological account. In *Voicing in Japanese*, ed. Jeroen van de Weijer, Kensuke Nanjo, and Tetsuo Nishihara, 157–176. Berlin and New York: Mouton de Gruyter.

Zuraw, Kie. 2000. Patterned exceptions in phonology. Doctoral Dissertation, University of California, Los Angeles.

Zuraw, Kie, and Sharon Peperkamp. 2015. Aspiration and the gradient structure of English prefixed words. In *Proceedings of the International Congress of Phonetic Sciences*, ed. The Scottish Consortium for ICPhS 2015, volume 18, 0382: 1–5.